THE STRUCTURE AND DYNAMICS OF INFORMATION SHARING NETWORKS

by

Xiaolin Shi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Phyilosophy
(Computer Science and Engineering)
in The University of Michigan
2009

Doctoral Committee:

Assistant Professor Lada A Adamic, Co-Chair
Associate Professor Martin J Strauss, Co-Chair
Professor Hosagrahar V Jagadish
Associate Professor Kevin J Compton
Associate Professor Anna C Gilbert
Associate Professor Dragomir R Radev

To Mom and Dad

ACKNOWLEDGEMENTS

First of all, I would like to thank my research advisor, Prof. Lada Adamic. This thesis would not be possible without her guidance, advice and encouragement. She sets a great model as a successful researcher for me to follow. I also would like to thank my academic advisor, Martin Strauss, for his support and advice, especially for those milestones during my PhD years, such as the prelim examination and the job search.

This thesis is the result of the joint endeavor with my collaborators, each of whom deserves my gratitude. Chapter II is joint work with Belle Tseng and Lada Adamic. Part of this work was done while I was a research intern at NEC Laboratories America in summer 2006. Chapter III is joint work with Matthew Borner, Lada Adamic and Anna Gilbert. Chapter IV is joint work with Lada Adamic and Martin Strauss. Chapter V is joint work with Lada Adamic, Belle Tseng and Gavin Clarkson, and part of this work was done while I was a research intern at NEC in summer 2007. Chapter VI is joint work with my colleagues, Jun Zhu, Rui Cai and Lei Zhang, when I was a research intern at Microsoft Research Asia in summer 2008. All the collaborating experiences were wonderful and fruitful. I would especially like to thank Dr. Belle Tseng, who was my mentor at NEC Laboratories America during the summers of 2006 and 2007. She provided me with great freedom in pursuing research topics of my interests. I would thank Belle and other members in her group at NEC: Xiaodan Song, Yun Chi and Koji Hino, for their countless help, research

insights and interesting discussions. The summer I spent at Microsoft Research was another terrific experience. Jun Zhu was a great collaborator and supportive friend. Rui Cai and Lei Zhang all nicely provided me with lots of help and conveniences for my work.

I would like to acknowledge my thesis committee members, Prof. Kevin Compton, Prof. Anna Gilbert, Prof. H V Jagadish and Prof. Dragomir Radev. Their insightful comments and kind support have improved this thesis a lot.

Another professor I would like to acknowledge is Prof. Yaoyun Shi. He served as my advisor in my first year, and it was him who led me into the graduate life at the University of Michigan.

My graduate life would have been much tougher without my wonderful friends at Michigan: Bin Liu, Joseph Xu, Ying Zhang, Yunyao Li, Hailing Cheng, and many others. Their warmest cares were the best support for me during the long, cold winters.

Finally, I would thank my parents. Their eternal and selfless love is the best motivation for me to pursue my dream bravely and consistently.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

The importance of information sharing networks is gaining increasing attention from research scientists. These networks play a crucial role in how we acquire information, how we convey information to one another, and how we interact with other people. Many information sharing networks can help normal users with various daily activities, such as reading and recommending news, making or contacting friends and online purchasing. All of those activities are actual electronic transactions and can be recorded. Nowadays, such data are continuously growing and evolving, and are an indispensable source of information for researchers to study the underlying human dynamics that are reflected by its various patterns. Most of such data has either explicit or implicit link patterns, and these links are potential paths for information to spread over that entire online social environment.

Due to the wide usage of online social environments in daily lives and the availability of the data, this type of data helps researchers investigate the behavioral patterns of people foraging for information and interacting with each other in two ways [53]. First, as such networks are of unprecedented size and are evolving rapidly, they are increasingly closer to the human activity dynamics of the real world. Second, information sharing networks are prevalent and important in people's daily activities so

that human behavior is greatly affected by them. For example, networks of emails or online social networks are two of the major ways that people communicate and socialize with others. As these two ways converge, the research that analyzes and studies information sharing networks reveals more truths about real-world human behavior and their dynamic systems.

Most research work that has been done on information sharing networks can be mainly classified into three categories. The first type of research is done by defining networks from real online data and studying what properties the networks have, either static or with temporal changes [81]. It is well known that, in spite of the random dynamic changes of information systems, there are a number of strong regularities both in the structural and temporal features of those networks, such as the power-law of the degree distribution and the small world phenomenon of the Web graph [2, 23]. The second is understanding why networks have those properties. Many features of the networks are explained and simulated by some relatively simple dynamical mechanisms or modeled by some simple rules in random graphs, such as the preferential attachment mechanism [15] and the forest fire models [68]. These models help the understanding of online networks and thus to predict their future behavior. The third is utilizing the properties of real-world networks to achieve some tasks, such as searching for information [86], summarizing the data [104] and finding communities [87]. Such research helps the development of more powerful and efficient algorithms and software for collecting and retrieving information, and it also provides valuable insights for designing better systems for users. The research work in this thesis, which is about the structural features of information networks and their underlying relationship with information dynamics, is mainly in the first two categories. However, many results also imply potential applications in information

management, information retrieval, etc.

As this thesis focuses on uncovering the features of structure and dynamics of information sharing networks, it studies the topological structures of these networks, the relationship of the structures and information diffusion, and the influences of information flows on network evolution. The first three pieces of work (Chapters II, III and IV) aim to understand the structures of networks upon which information flows, and the last two pieces (Chapters V and VI) are trying to answer the question of how the network structure influences information diffusion and the relationship between communities and information diffusion.

Before looking at the structural features of information sharing networks in detail, we should be aware that such networks are massive and rapidly evolving, so that their properties are difficult to trace. Thus, the first question we have is whether the structure resulting from information sharing can be reliably measured in networks that are continually evolving. In Chapter II, we answer this question by studying sampled data sets of *blogosphere*. The blogosphere serves as a medium for self-expression, community formation and communication, and information diffusion and aggregation. The rich structure of the blogosphere has proven to be fertile ground for exploring research questions from a variety of fields. Some have focused on the motivations behind blogging [21], the relationship between a person's tendency to keep blogging and their embeddedness in the online social network [63], and explored the possibility of extending blog's interactive nature for research and commercial collaboration [20]. A few studies have specifically focused on the LiveJournal blog network and found patterns of link distribution across geography [70], factors contributing to link formation such as common interests and age [59], and even the likelihood of a blogger joining a new LiveJournal interest group if many of

their blogging friends have [13]. Others, closer to our current goals, have pursued a systematic approach of analyzing the large scale network structure of the blogosphere. Kumar et al. examined the structure of the blogosphere, both in terms of the bursty nature of linking activity, the uneven distribution in the concentration of such links, and the effect of time windowing on the appearance of that distribution [58, 59, 60]. Information diffusion studies have aimed to use the link structure and other blog properties to infer the path of information flow [45, 4]. Moreover, other studies have used the link structure to solve problems such as splog detection [56] and community identification [117]. Splogs (also known as spam blogs) are blogs whose sole purpose is to direct traffic and increase the search engine rankings of particular websites. In this work, by comparing two large blog datasets, we demonstrate that samples from the blogosphere might differ significantly in their coverage but still show consistency in their aggregate network properties. The results of the work also show that properties such as degree distributions and clustering coefficients depend on the time frame over which the network is aggregated [101].

While Chapter II shows that it is possible to get reliable metrics of real-world networks based on comprehensive samples, we may still face challenges in obtaining such samples, when, for example, the network is simply too massive. To address this problem, in Chapter III, we observe that some vertices in many large information sharing networks can play an important role in graph representation and information diffusion. The numbers of such special vertices can be very small compared to the size of the original networks. In this chapter, which deals with *important vertices and their graph synopses*, we examine the properties of subgraphs of the most prestigious vertices, i.e., vertices of the highest values using some well-established importance measures, in several online networks, including those of blogs, websites in general, and

instant messaging users. There are previous studies about compressing web graphs for space-efficient data storage and transfer [116, 5], using a subgraph to represent the original large graph (the graph sampling problem) [67, 62], mining a subgraph for visualization of the original graph [39, 123], placing sensors to detect information flow [65], constructing a synopsis by projecting queries [66], and quantifying the extent to which important vertices hold online social networks together [76]. However, this work has focused on keeping or representing the properties of the original networks; i.e., studying the entire networks. We study the more fundamental properties of the subgraphs induced by important vertices. Our principled and rigorous study of the properties, construction and utilization of subsets of special vertices in large online networks showed that vertex-importance graph synopses provide small, relatively accurate portraits, independent of the importance measure [104].

In addition to revealing interesting characteristics of the vertices in information sharing networks and their roles in information diffusion, further investigation about a set of special edges, the *strong ties*, and their relationship with information diffusion are studied in Chapter IV. Simple connectivity through arbitrary ties is sometimes insufficient to transmit information, because ties may need to be of a given strength in many real-world scenarios. For example, sensitive information or information that may confer an advantage to those who have it, may only be shared between individuals who know one another well [46]. This chapter analyzes the connectivity and information transmission of strong ties. From some online social networks, we show that strong ties occupy a large portion of the network and that removing all other ties does not change the majority of the giant connected component and the average shortest path of the online friendship networks. What is more, the cost of forming transitive ties (which we take as the definition of strong ties) by modeling

a random graph composed entirely of closed triads is evaluated. Both the empirical study and random model point to the robustness of strong ties with respect to the connectivity and small world properties of social networks. Thus, this work shows that it is still possible, under the restriction of tie strength, for information to be transmitted widely and rapidly in empirically observed social networks [102].

After examining the structural features of the networks in which the information is possibly transmitted, this thesis further investigates how the structure influences information diffusion and the relationship between network communities and information diffusion. In Chapter V, we examine information diffusion between communities and its subsequent impact in information sharing networks by studying the *citation networks*. Published scholarly work is a traditional social medium for the exchange of scientific ideas and knowledge. The structure and growth of such citation networks have been studied extensively to measure the impact of individual articles and the evolution of entire fields [29, 31, 93, 17]. Applying bibliometrics to citation networks to study the impact of fields, individuals, and particular papers has been the purview of the field of scientometrics [31]. As early as in the 1960s, de Solla Price first developed models to explain the heavy tailed distribution in the citations an individual paper receives [29]. Recently, the availability of large scale citation data and computational power has enabled the visualization and quantification of the amount of information flow between different areas in science [19, 16], in effect mapping human scientific knowledge. These visual maps leave open the question, however, of the size, speed and impact of information flows across community boundaries. Prior work has shown these flows to be relatively insignificant; omitting information flow between communities when one models citation networks still provides realistic citation distributions and clustering coefficients [17, 97]. Not only are information flows

across scholarly communities infrequent, they are also delayed: on average more time elapses between the citing and cited articles for citations across disciplines than ones within a discipline [93]. In this chapter, we view citation networks from the perspective of information diffusion. We study the structural features of the information paths through the networks and analyze the impact of various citation choices on the subsequent impact of the publication. The analysis shows that a publication's citing across disciplines is tied to its subsequent impact. In the case of patents and natural science publications, those that are cited at least once are cited slightly more when they draw on research outside of their area. In contrast, in the social sciences, citing within one's own field tends to be positively correlated with impact [100, 103].

In Chapter VI, we study the factors that could influence information diffusion among online communities. We examine the diffusion curves and the likelihood that a user will join a group based on the pattern of her interaction with other users and the features of groups in online forums. As new ideas and controversial discussions are always emerging and propagating among online forums, it is interesting to study the process of information diffusion in this social medium. The human behavior of gathering together and forming groups has been an important theme in studying information diffusion, because people taking the same actions as their neighbors is strong evidence that information flow has occurred [13]. Characterizing user grouping behavior in online social environments does not only help researchers to understand many of the sociological problems of human behavior, but also facilitates them to improve various applications in the online environment, such as the recommendation systems [110]. In this chapter, patterns in user behavior in joining groups and the feature factors associated with users or groups that influence such behavior are studied. We show the diffusion patterns associated with features of

users and groups, and we use Markov random graph models to help understand the relationships of these features, as well as the differences in their impact in different types of online forums [105].

At last, we will conclude the work in this thesis, and discuss future work in perspective at Chapter VII.

CHAPTER II

Sampled Data of Blogosphere

2.1 Introduction

In this chapter we first address the question of how information sharing networks, which are constantly evolving, can be captured and understood. Blogs are especially well suited for this study, since they form a vast dynamic and growing network, with new blogs continuously emerging, millions of existing blogs creating new content, while some lay abandoned as their authors start other blogs or activities. Of particular interest are the direct citation patterns between the blogs, because they indicate interaction and information diffusion—blogs linking to posts they read on another blog while possibly writing additional content of their own. Tracking information diffusion in the blogosphere is not just an intriguing research problem, but is of interest to those tracking trends and sentiments. Several online services, such as BlogPulse and Technorati, report the most actively discussed topics in the blogosphere. A heavily blogged topic, even if it originates in the blogosphere, is likely to make its way into the mainstream media. In fact, many mainstream media sources now host blogs as an integral part of their websites, while some of the most popular blogs rival most mainstream media online outlets in popularity [10].

In this chapter, we have two objectives. The first is to examine how robust the

features of the blogosphere are when examined through the lenses of two different samples. The second is to compare these features with previously studied Web and social network datasets, in order to understand the blogosphere network structure in the wider context of other social and technological networks. Our blog datasets stem from two sources, BlogPulse and TREC (described in more detail below), both intended for use by the research community to study different aspects of the blogosphere. We compare these two sets of blogs directly, first in terms of their coverage and overlap, then in terms of their network properties.

We find that although the datasets differ widely in size, cover different time durations, and are set months apart, their properties show remarkable consistencies. Unfortunately, a fair fraction of blogs are in fact spam blogs, automatically generated blogs created with the intention of altering search engine results and directing traffic to specific websites. These splogs account for a large fraction of the links in the datasets. Consequently, we also study the effect of splog removal on the properties of the networks. Furthermore, we examine the effect of aggregating the network over time, similar to previous work by Kumar et al. [60], and find that the degree distribution and other properties converge when the network is aggregated. Finally, we contrast the linking patterns within and between different blog hosting sites, finding that most large blog hosting sites tend to be "exporters" of links—with many of those links going to blogs with their own domain names.

2.2 Description of data sets

We use two datasets in our study of the blogosphere. One is the WWW2006 Weblog Workshop dataset from BlogPulse, which has 1,426,954 blog URLs in total, and 1,176,663 distinct blog-to-blog hyperlinks. This dataset covers 3 weeks of blogging

activity, from July 4, 2005 to July 24, 2005. It contains hyperlinks that occur only in the blog entries themselves, and exclude blogrolls or comments. Consequently, the network is quite sparse—among the over 1.4 million blogs, only around 141,046 (10%) of them have links to other blogs in the dataset. If we only consider the blogs having at least one in-link (receiving a citation from another blog) or out-link (giving a citation to another blog) in the dataset, the average degree of this network is $\langle k \rangle = 4.924$. We omit from our analysis the additional 160,670 URLs, that were identified by BlogPulse to be blogs with at least 1 citation, but whose entries were not included in the dataset.

The TREC (Text REtrieval Conference) Blog-Track 2006 dataset is a crawl of 100,649 RSS and Atom feeds collected over 11 weeks, from December 6, 2005 to February 21, 2006. In our experiments, we removed duplicate feeds and feeds without a homepage or permalinks. We also removed over 300 Technorati tags (e.g., Technorati.com/tag/war_on_terror), which appear to be blogs, but are in fact automatically generated from tagged posts. Different from the BlogPulse data, the TREC dataset contains hyperlinks of various forms, including blogrolls, comments, trackbacks, etc. There are 198,141 blog-to-blog hyperlinks in total, and 33,385 blogs having at least one such link. However, in order to do a fair comparison of the two different blog datasets, we restrict the TREC data to only the 61,716 hyperlinks occurring within entries. There are 16,432 making or receiving at least one such link, giving us an average degree of $\langle k \rangle = 3.8$. The work of [73] describes the creation of the TREC data in more details, and reports some statistics about this dataset, such as the degree distribution.

Aside from the differences in the sizes and time spans of the two blog datasets, the nature of the two corpora and the way they are constructed are also different.

The BlogPulse dataset is more like a complete snapshot, while the TREC dataset is more biased and artificially sampled. Considering all these factors, one of the main purposes of our work is to explore how these factors would affect the observations of blogosphere.

Some previous work has identified a certain fraction of splogs in these two datasets. In BlogPulse, according to the splog detection methodology presented in [56], the percentage of splogs is 7.48%. And in TREC, the percentage of splogs is about 18% [73], while after restricting the blogs to those have homepages, the percentage of splogs detected was around 7% [71].

2.2.1 Dataset overlap

The BlogPulse and TREC datasets are two samples of the same blogosphere, albeit of vastly different sizes, covering different time durations and about 5 months apart. We are interested in comparing them in two respects in order to assess the difficulty in obtaining a comprehensive sample of the blogosphere. First, we compare the coverage of the two sets according to different blog hosting sites, as shown in Figure 2.1. In the TREC dataset, a smaller fraction of the blogs is hosted by the major blog hosting sites. The largest subset at 28% is hosted by LiveJournal, followed by 6% hosted at TypePad. In contrast in the BlogPulse dataset , a full 48% is hosted at LiveJournal, followed by 20% hosted at Xanga.

Second, we directly compare the overlap between the two blog datasets in terms of the blogs commonly crawled by both. Figure 2.2(a) shows that of the 16,432 blogs whose entries were included in the TREC dataset, 7,225 (or 44%) are also in the much larger BlogPulse dataset. Finally, we take this common set of blogs and compare the overlap in the undirected edges in the two datasets. Specifically, if blog A cites blog B (or vice versa) during the 3 week period covered by the BlogPulse

data, we examine what percentage of the time we also observe blog A citing blog B or vice versa in the 11 week period of the TREC crawl 5 months later. Somewhat surprisingly, we find very little overlap. There were only 2823 pairs with edges in both datasets, compared to 56,387 pairs with edges in the BlogPulse data and 57,091 in the TREC data. This means that the same blogs that might be mentioning one another during one short period have only a 5% chance of doing so about half a year later. The above shows us that two relatively large datasets representing "samples" of the blogosphere actually have dramatically different coverage of blogs. Even where the two network samples overlap in nodes, we find that the connectivity, namely the links between the blogs, are likely to change substantially over time.

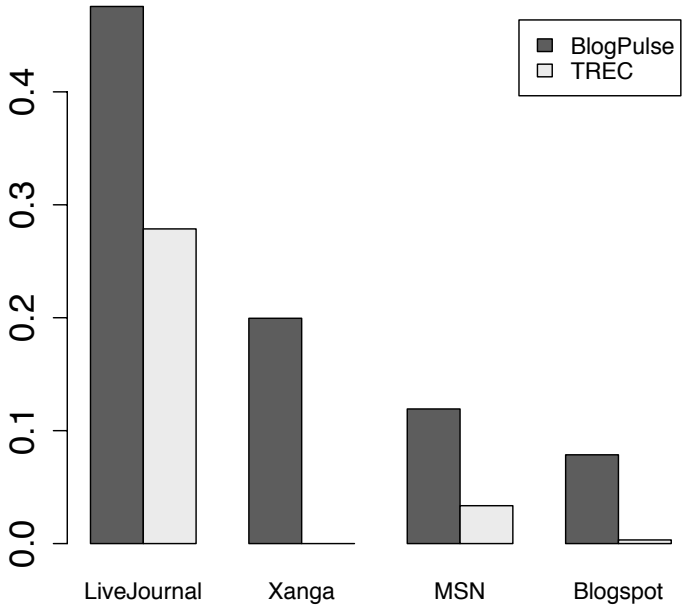


Figure 2.1: Proportion of blogs at 4 large blog hosting sites over the two datasets, demonstrating that TREC is less concentrated at large hosting sites than BlogPulse

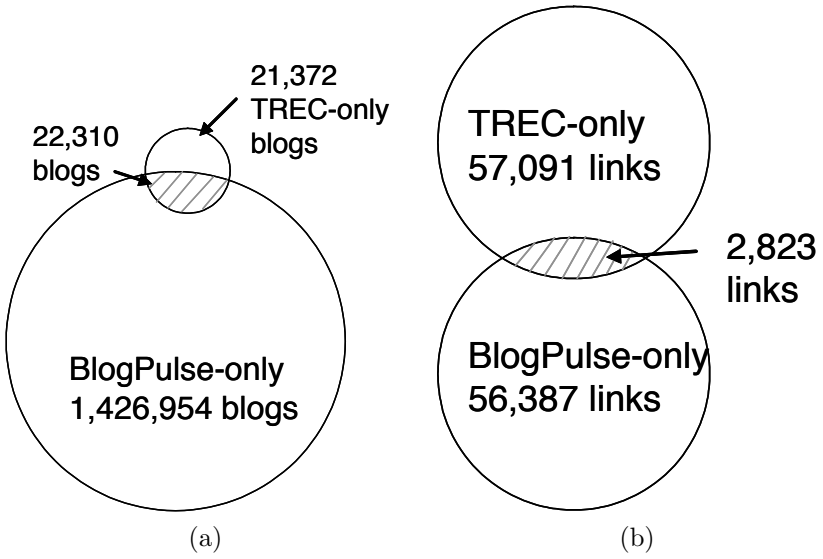


Figure 2.2: Overlap in coverage between TREC and BlogPulse: (a) overlap in crawled blogs with around 50% of the blogs covered in TREC being covered in the BlogPulse sample (b) overlap in between-blog links in the TREC and BlogPulse datasets restricted to blogs that occur in both

2.2.2 Other network datasets

To understand the properties of the blogosphere and how they differ from other networks, we study similar features in the Web graph, which was presented in [8] and [23]. The former dataset contains 325,729 documents and 1,469,680 links taken from a 1999 crawl of the `nd.edu` domain. The latter crawl from 2000 contains 200 million web pages and 1.5 billion links.

2.3 Topological features and network comparisons

In this section, we study the properties and topological features of the blogosphere by analyzing the two networks constructed from the BlogPulse and TREC datasets. We first restrict our analysis to those links that are located within crawled entries and cite blogs within the data set. We then include any additional hyperlinks, such as blogrolls, comments, and trackbacks, that were included in the TREC dataset. The networks are treated as directed but unweighted graphs where we are simply

taking into account whether a blog cites another blog, and not how many times it does so.

2.3.1 Degree distributions

In a directed graph, for a vertex v, we denote the *in-degree* $d_{in}(v)$ as the number of arcs to v and the *out-degree* $d_{out}(v)$ as the number of arcs from it. The distribution of in-degree p_{in}^k is the fraction of vertices in the graph having in-degree k and p_{out}^k is the fraction of vertices having out-degree k. If both the in-degree and out-degree of a vertex are 0, then the vertex is *isolated*.

The *average in-degree* is:

$$\langle k \rangle_{in} = \frac{1}{|V|} \sum_{v \in V} d_{in}(v) \tag{2.1}$$

which is a global quantity but measured locally. The *average out-degree* $\langle k \rangle_{out}$ is defined similarly.

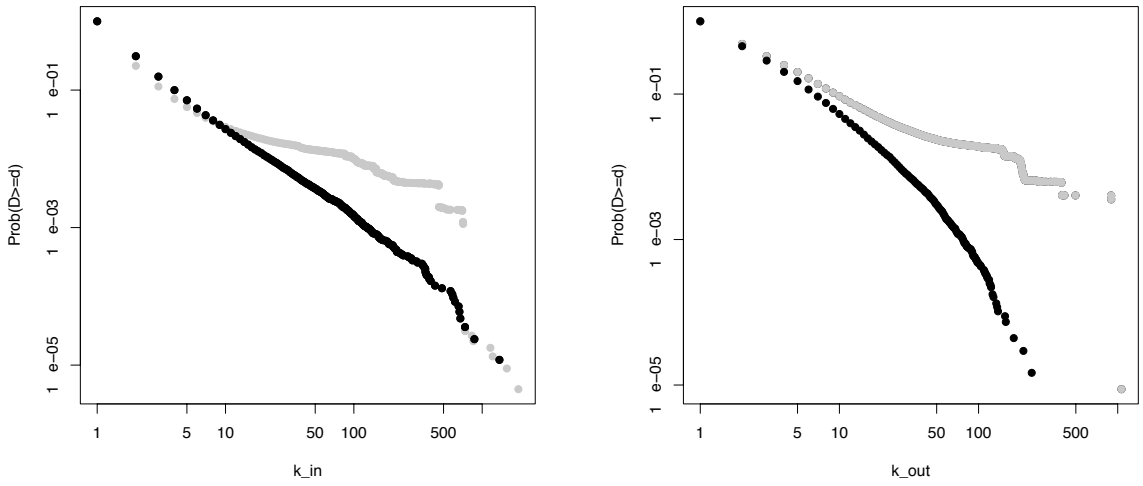


Figure 2.3: The degree distributions of the BlogPulse data with splogs (grey curves) and without splogs (black curves)

First, we observe that the degree distributions are greatly affected by the existence of splogs. Considering all the blogs in the BlogPulse data, both in-degree and out-degree distributions have an unusually high number of blogs with degrees ranging from 10 to 500. This results in irregular shapes for the cumulative degree distributions, which represent the proportion of blogs having at least k in-links or out-links. However, after removing splogs identified by Pranam et al. [56] for the BlogPulse dataset, we replicate the result that the cumulative in-degree and out-degree distributions show smoother curves, as shown in Figure 2.3.

After excluding splogs from the BlogPulse data, we compare the degree distributions of the blogosphere and the Web, using the Web degree distributions measured by Broder et al. [23] for a 1999 Alta Vista crawl of 200M pages. This previous study, along with a study of the `nd.edu` domain [8, 7], and a crawl in 2001 of 200M pages by the WebBase project at Stanford [32] found that the indegree distribution of the Web is scale free with a power law exponent α of 2.1. From Figure 2.4, we can see that the in-degrees of the BlogPulse and TREC datasets show similar power-law distributions to the Web graph. TREC exhibits a slightly shallower slope, while the BlogPulse data presents a slightly steeper one. This is consistent with the previous finding that sampling a power-law network can produce networks with steeper power laws [115]. Since the BlogPulse data is of a shorter time duration than the TREC data it may be more likely to resemble a subsample of the full network. Of course it is difficult to directly compare a web page crawl which contains a single static snapshot of a page, with an aggregation over 11 weeks of an RSS feed for a blog. Although a single download of a blog would usually contain a limited number of entries (with previous ones usually moved to an archive), the RSS feed would correspond to a single long page where content is added over time, but not deleted. It is possible

that this aggregation over a longer time period accounts for the similarity of slope for the TREC data compared to the Web.

The Web outdegree distribution has been found either not to follow a power law distribution at all, or to exhibit a steeper power law only in the tail. Broder et al. measured the tail to have a power-law exponent of 2.7, Albert and Barabasi measured it to be 2.45 [8, 7], while Donato et al. found it not to follow a power-law at all [32]. The out-degree distributions of TREC and BlogPulse, shown in Figure 2.5, drop off much more rapidly than the Web graph. On the one hand, this may again be due to sampling. For example, Pennock et al. [90] showed that when certain subcategories of pages are sampled, what starts out as a power-law degree distribution can exhibit sharp drop-offs. Certainly, blogs are only a subcategory of all web pages, and we are furthermore only considering links among a sampled set of blogs. But, more likely, the number of hyperlinks a blog can generate in a limited time period is bounded. This constraint is also observed in many social networks, e.g., co-authorship networks [79]. So while it is possible for one blog to gather much attention (inlinks) in a short time period, it appears less likely for a single blog to lavish as much attention on as many different blogs in the same time period. The same tends to hold true on the web, where some webpages are linked to by thousands of others, but it is much less likely for a single page to contain thousands of hyperlinks.

Our results also concur with previous measurements of the blogosphere, which have revealed power-law distributions of in-degrees based on blogrolls and in-post citations [60, 70, 106]. Here we were interested in whether we would still observe the power-laws when considering only the in-post citations.

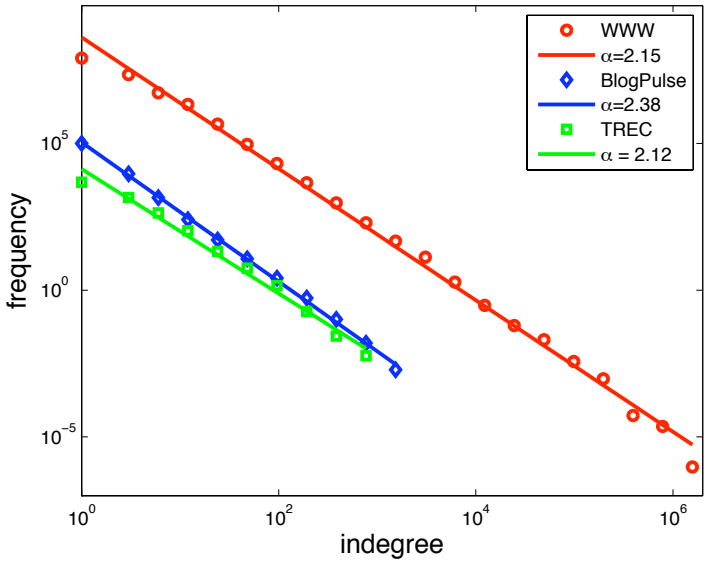


Figure 2.4: In-degree distributions of Web, BlogPulse and TREC data, exponentially binned, all showing power-law structure

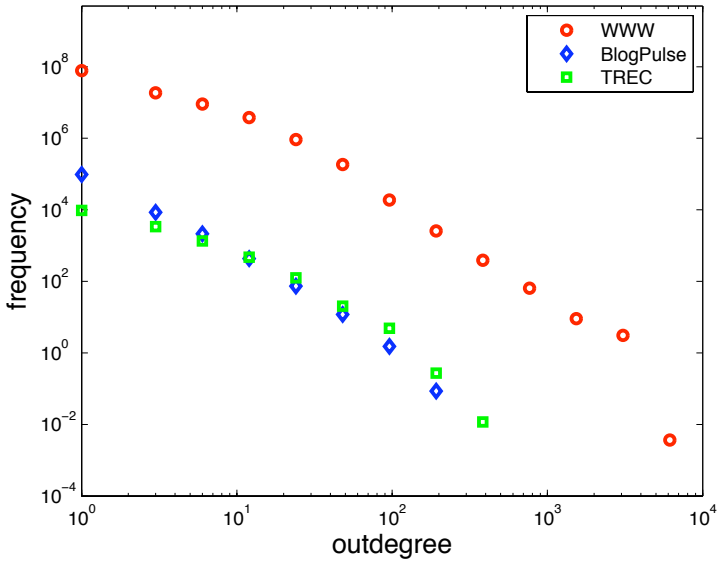


Figure 2.5: Out-degree distributions of BlogPulse, TREC and Web, exponentially binned

2.3.2 Small-world effect

The small world effect states that in the network, the average shortest path between every pair of reachable nodes is short compared to the total number of nodes in the network[52].

The studies of Web graph have shown that the WWW has the small-world property. Even as the number of web pages has grown exponentially, the average number of hyperlinks that need to be traversed to get from one page to any other (provided such a path exists) has remained relatively small. Albert et al. [8] give a formula to compute the average shortest paths in Web graph if the number of web pages N is known:

$$\langle l \rangle = 0.35 + 2.06 \log(N) \tag{2.2}$$

The estimate for the average shortest path using this formula for a graph of 200 million nodes is $\langle l \rangle = 17.45$, which is quite close to the measured value ($\langle l \rangle = 16$) found by Broder et al. [23] for a data set of 200 million web pages.

Quite similar to the Web graph, our experiments show that even when considering only entry-to-blog links, the blogosphere has the small-world property. Our two datasets, although of different time durations and only partial overlapping in blogs and links, have very consistent shortest paths considering their network sizes. For the TREC dataset (16,432 blogs), it is $\langle l \rangle = 7.12$, and for the BlogPulse dataset, which is of 143,736 blogs, it is $\langle l \rangle = 9.27$. If we let $N = 16,432$ or $N = 143,736$ in Formula 2.2, then the $\langle l \rangle$s calculated for TREC and BlogPulse are 9.03 and 10.97 respectively, which are larger than what they have been in our experiments.

However, this does not necessarily mean that it is easy for information to diffuse

widely in the blogosphere. This is because information diffusion is not only related to the average shortest path, but also the connectivity of the graph. Since the average shortest path is only computed between all reachable pairs, it doesn't take into account what proportion of pairs of blogs could not be reached one from the other simply by following hyperlinks. Our experiments show that only 12.37% of the pairs of blogs in TREC are reachable. For the BlogPulse data, only 6.13% of the pairs of blogs are reachable. Even when we consider the network of TREC data with all forms of hyperlinks contributing to its edges, the percentage of reachable pairs is still only 22.11%. The low percentage of reachable pairs of nodes is also true in the Web: over 75% of time there is no directed path from a random start node to a random destination node [23]. If the connectivity is low in the network, as it is for the TREC and BlogPulse data, it will yield a small average shortest path, but at the same time produce many infinite paths that are not counted.

In the following section we examine the important question of connectivity in more detail.

2.3.3 Connectivity

For a directed graph, there are two types of connected components: *weakly connected components* (WCCs) and *strongly connected components* (SCCs). A strongly (weakly) connected component is the maximal subgraph of a directed graph such that for every pair of vertices in the subgraph, there is an directed (undirected) path from v_x to v_y. Thus a weakly connected component is a larger subgraph than a strongly connected component.

In the two blog datasets, within a weakly connected component one can follow links within posts to reach either blog A from blog B or vice versa (but not necessarily in both directions), for each pair of blogs A and B. In practice these paths

Network	# of nodes	Max WCC	Max SCC	Fraction of SCC in WCC
Web [8]	325,729	325,729 (100%)	53,968 (16.57%)	16.57%
Web [23]	203,549,046	186,771,290 (91.76%)	56,463,993 (27.74%)	30.23%
BlogPulse	143,736	107,916 (75.08%)	13,393 (9.32%)	12.41%
TREC	16,432	15,321 (93.24%)	2,327 (14.16%)	15.19%

Table 2.1: Connectivity comparison between the Web graph and blogosphere samples

may be hard to find because the link leading to the path to the second blog could be in any one of the posts made over the 3 or 11 week period. Nevertheless, the connected components give us a sense of the connectedness of the datasets. Our experiments show that, in the TREC data, the largest weakly connected component includes 15,321 nodes, and the largest strongly connected component is of size 2,327. The sizes of largest weakly connected component and strongly connected component of BlogPulse data are 107,916 and 13,393 respectively. We have a comprehensive comparison of the connectivities of blogosphere and the Web in table 2.1. Similar to the Web [33], the discrepancies in the size of connected components are most likely due to the different ways the datasets are crawled, and the time periods in which the networks form. In section 4, we will further discuss the temporal features of the connectivity of blogosphere.

On the other hand, if we also consider other forms of hyperlinks in TREC, including 33,385 blogs and 198,141 blog-to-blog hyperlinks, then the resulting network has much better connectivity. The size of the largest weakly connected component is 88.93%, and the size of largest strongly connected component is 44.36%. This shows that the blogosphere is glued together by blogrolls, even if over a limited time period there is relatively little active citation.

Another interesting observation about the connectivity of the blogosphere is the following: before cleaning the TREC data, there are 363 **technorati.com** tag URLs, with 47,521 links either from or to these URLs. Our experiments show that the

existence of such extremely high in-degree or out-degree nodes does not affect the overall connectivity of the blogosphere. Before removing the Technorati tag URLs and their links, the size of largest weakly connected component is 30,180 (90.40% of the whole network) and the size of largest strongly connected component is 15,176 (45.46%) - only slightly larger than the components with the tag URLs removed. This observation is similar to the one made for the Web graph by Broder et al. [23], showing that high degree nodes do not play the function of "junctions" in the connectivity of the Web.

2.3.4 Clustering coefficient and reciprocity

The *Clustering coefficient* is a measurement of the percentage of closed triads in a network. For every vertex v_i, its clustering is defined as:

$$C_i = \frac{\text{number of closed triads connected to } v_i}{\text{number of triples of vertices centered on } v_i} \tag{2.3}$$

Then the clustering coefficient for the whole graph is averaged over all vertices i.

In an Erdös-Renyi random graph (a random graph in which every pair of vertices are connected by probability p) [35] with n nodes and a constant average degree, the clustering coefficient is $O(n^{-1})$. However, in most real-world networks, the clustering coefficient, $O(1)$, is much higher, reflects the prevalence of closed triads [81]; i.e., if vertex v_x is connected to vertices v_y and v_z, then the probability for v_y and v_z to be connected is higher than expected at random. For measuring the clustering coefficient in a directed graph, we ignore the directions of arcs.

The clustering coefficients of TREC 0.0617 and BlogPulse 0.0632 (including splogs in both datasets) are large compared with what they are in the corresponding Erdös-Renyi random graphs. We see that for these values of clustering coefficients, the two

datasets are showing nice consistency in spite of the differences in crawling and time duration. These high values are also similar to measurements of the clustering for the Web graph ($C = 0.29$ [81]) and co-authorship networks ($C = 0.19$ [79]).

Reciprocity is another measurement that shows a significant difference between real-world networks and the Erdös-Renyi graph. The reciprocity values (how often, when A links to B, B links to A) is another measure of cohesion, reflecting mutual awareness at a minimum, and potentially online interaction and dialogue. In the datasets, we actually observe very little reciprocity: in TREC, 4.98% edges are reciprocal, and in BlogPulse 3.29% edges are reciprocal.

However, if we also consider other types of links in TREC, making the network significantly denser, then the clustering coefficient of this graph is 0.13, and the reciprocity is 20.06%, both of them are significantly larger than they are in the blogosphere merely with entry-to-blog links as its edges. A possible explanation is that people often create entry-to-blog links to cite information. Other types of links, such as comments and trackbacks are by their nature interactive (and trackbacks are by definition reciprocal). Even blog rolls may exhibit higher reciprocity, because bloggers tend to list their friends' blogs as well as other blogs they tend to read, and friendship is often, though not always, reciprocal. Therefore the low reciprocity we observe could be due to the nature of entry-to-blog links themselves and the short time window of the samples, where we simply haven't waited long enough to observe a reciprocal link.

2.4 Temporal features

As we have described before, the time ranges of the two datasets are of different lengths: the BlogPulse sample covers 3 weeks, while TREC is crawled over 11 weeks.

In order to explore the effects of the crawling periods on the observations of the blog datasets, we take the longer-period TREC dataset and study the properties of the subgraphs in the TREC network over 4 different time windows.

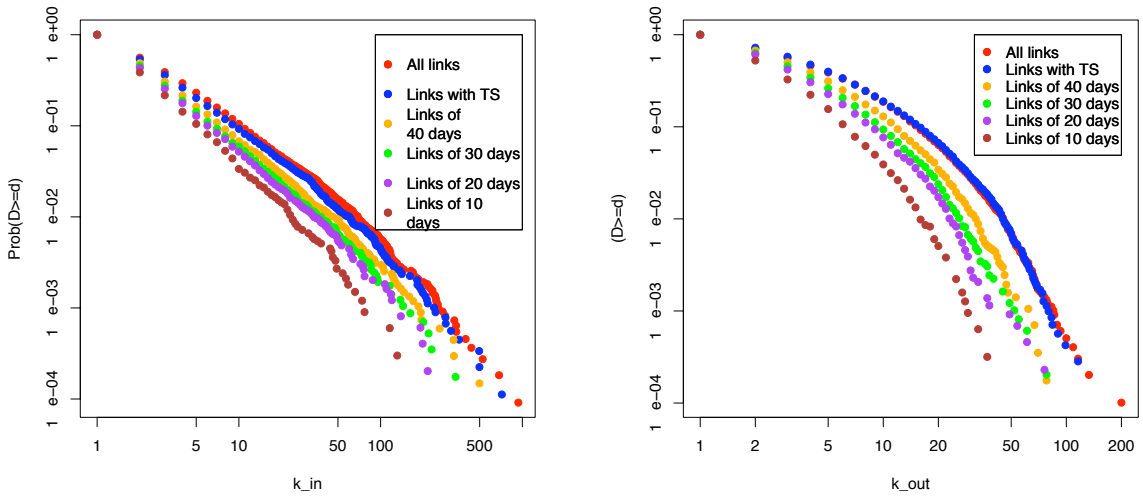
We assign a timestamp for each entry-to-blog link as the time the entry is created, where 74.72% of the entries have timestamp. Four overlapping time periods are chosen corresponding to the first 10, 20, 30 and 40 days of the TREC crawl. The 10 days capture 5,793 blogs and 8,818 entry-to-blog links with an average degree of $\langle k \rangle = 1.5$. The first 20 days capture 8,054 blogs, 16,206 links, bringing the average degree up to $\langle k \rangle = 2.0$. The first 30 days capture 9,085 blogs with 20,411 links and $\langle k \rangle = 2.2$. The last subset of 40 days contains 10,433 blogs and 27,724 links with $\langle k \rangle = 2.657$. This illustrates that as the time duration increases, the average degree also increases.

2.4.1 Degree distributions

We plot the degree distributions of the four time-overlapping subnetworks (10 days, 20 days, 30 days, 40 days), as well as the entire network of 11 weeks with link time stamps (denoted by "Links with TS"), and the entire network with or without link time stamps (denoted by "All links"). From the in-degree and out-degree distributions in Figures 2.6, it is apparent that different time windows yield very similarly shaped curves for both the indegree and outdegree distributions. However, as the time periods get shorter, the curves for both in-degrees and out-degrees are steeper.

2.4.2 Connectivity

In section 3.3, we found that both the BlogPulse and TREC samples have large weakly connected components, but relatively small strongly connected components,



(a) In-degree distributions

(b) Out-degree distributions

Figure 2.6: Temporal changes in the in-degree and out-degree distributions in TREC

Subset	# of nodes	Max WCC	Max SCC	Fraction of SCC in WCC
First 10 days	5,793	4,719 (81.46%)	None (0%)	0%
First 20 days	8,054	7,162 (88.92%)	349 (4.33%)	4.87%
First 30 days	9,085	8,249 (90.80%)	471 (5.18%)	5.71%
First 40 days	10,433	9,662 (92.61%)	730 (7.00%)	7.55%
All blogs	16,432	15,321 (93.24%)	2,327 (14.16%)	15.19%

Table 2.2: Temporal changes in the connectivity in TREC

with the TREC sample showing better connectivity than BlogPulse, in spite of BlogPulse having larger average degree. The dynamics in the connectivity of blog subgraphs over different time windows is shown in Table 2.2. As time goes on, both the size of the largest weakly connected component and the size of the largest strongly connected component grow larger, and thus the connectivity is increasing. It can also be observed that the weakly connected component is formed earlier and grows more rapidly. In contrast, it takes a much longer period for the strongly connected component to form; however, after a certain period of time, the growth of the largest weakly connected component is relatively stable near 100% of the network, while the largest strongly connected component continues to grow.

2.4.3 Clustering coefficient and reciprocity

Next we examine the temporal changes in reciprocity and the clustering coefficient. Our experiments show that the values of reciprocity of the links from the first 10 days to the first 40 days are 2.88%, 3.85%, 3.84% and 4.12%. We can see that except for the shortest time period, all the other values are bigger than in the 3-weeks of BlogPulse (reciprocity of 3.29%) but smaller than in the entire 11-weeks of TREC (reciprocity of 4.98%) . This indicates that reciprocity grows with time, because blogs have a longer opportunity to reciprocate. It also demonstrates that reciprocal links are still extremely sparse.

The clustering coefficients from the first 10 days to the first 40 days are 0.034, 0.043, 0.046 and 0.052. All of them are smaller than the clustering coefficient in both BlogPulse (0.0632) and TREC over the full time period (0.0617). So we know that although longer time would increase the clustering coefficient, it may depend more on the density of the sample.

2.4.4 Densification law

Leskovec et al. [68] described the densification law prevalent in many networks: the number of edges grows superlinearly in the number of nodes over time: $e(t) \propto n(t)^{\alpha}$. For example, in the Internet, there are new routers appearing and at the same time the number of connections per router is increasing, and the densification exponent is $\alpha = 1.18$. In patent networks, all the links are added from a patent at the time it is inserted into the network. The densification exponent is $\alpha = 1.66$. But in our network, probably most of the blogs already existed before the beginning of the crawl (it would be interesting to repeat the analysis with new blogs appearing). During the crawl, as more and more links are added into the network,

originally isolated blogs start connecting to each other. The number of edges, shown in Figure 2.7 is increasing nearly quadratically with the number of nodes ($\alpha = 1.928$). This relatively large value tells us that the densification of a crawled blogosphere with a static set of blogs is a faster process than some other networks such as the Internet and patent networks.

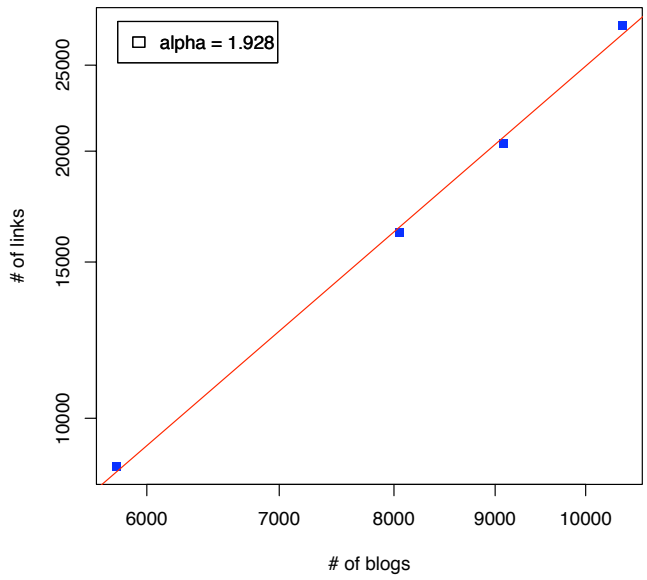


Figure 2.7: The number of edges versus number of nodes in log-log scale for blogs crawled over different time durations, which obeys the densification power law

2.5 Blogs in blog hosting sites

Another way of understanding the blogosphere is by analyzing it through different blog hosting sites. Currently, the four largest blog hosting sites are `LiveJournal`, `BlogSpot`, `Xanga` and `MSN`. They are also the largest four in the BlogPulse dataset, as shown in Table 2.3. In the table, "all links" either originate from or terminate at a blog at the specific blog hosting site; "in links" originate outside of the blog hosting site, but terminate within it; "out links" point from within the hosting site

to an outside blog; "internal links" lie between blogs within the hosting site. All these links occur only within entries, and no links of other forms, such as blogrolls, comments, etc. are included. The table also lists in italic the corresponding numbers of blogs and links after removing splogs according to [56]. An immediate observation we can make is that although splogs by number constitute only a small fraction of the total blogosphere, they account for a substantial proportion of the links.

# Blogs	All links	Inlinks	Outlinks	Internal links
LJ				
678,676	155,665	4,561	15,735	135,369(87.0%)
676,719	*95,161*	*2,718*	*8,731*	*83,712(88.0%)*
Xanga				
284,693	58,741	2,354	3,067	53,320(90.8%)
283,952	*8,454*	*437*	*1,534*	*6,483(76.7%)*
MSN				
170,108	58,811	44,180	1,528	13,103(22.3%)
162,147	*1,271*	*90*	*699*	*482(37.9%)*
BlogSpot				
112,184	845,093	34,979	73,730	736,384(87.1%)
62,256	*42,830*	*12,540*	*18,519*	*11,771(27.5%)*

Table 2.3: Blogs in hosting sites in the BlogPulse dataset

From Table 2.3, we also notice that, for most of these large blog hosting sites, no matter whether or not splogs are removed, the internal links usually occupy the greatest portion of the all links. The percentage of internal links of MSN is relatively small (22.3%). This is because in the BlogPulse dataset, there is a large portion of links from blogs of BlogSpot to blogs of MSN, which are mostly splogs (over 43,000 links). Due to this fact, it lowers the percentage of internal links of MSN. This is another aspect that tells us how splogs would affect our observations of blogosphere. This suggests that blogs within one hosting site are more likely to form densely connected communities, while it is less likely for blogs in different blog hosting sites to be in a community. This pattern may be a result of bloggers preferring to use the same hosting site as their friends, and different hosting sites being prevalent in

different countries. In the Table 2.4, we can see that links connecting two different blog hosting sites are very sparse, both with and without splog links.

Src & dst Blogs	LiveJournal	Xanga	MSN	BlogSpot
LJ	135,369 *83,712*	873 *159*	160 *10*	4,215 *1,714*
Xanga	1,208 *612*	53,320 *6,483*	124 *11*	659 *236*
MSN	61 *36*	179 *23*	13,103 *482*	309 *66*
BlogSpot	1,109 *707*	832 *151*	43,113 *17*	736,384 *11,771*

Table 2.4: Links among blog hosting sites in the BlogPulse dataset

Another thing one observes from the Table 2.3 is that the numbers of out links always exceed the numbers of in links for a blog hosting site, and most of those out links point to blogs with their own domain names. Since it is easier and often free to create blogs in the blog hosting sites, these kinds of blogs are more casual and personal; in contrast, blogs with their own domain names are more likely to be maintained in a more formal and professional way. And in this sense, it is natural for the self-hosting blogs to have more in links from other blogs.

2.6 Conclusions

For analyzing the topology of a large network such as the blogosphere, it is impossible for researchers to get all the data about it. Rather, one uses various sampling methods to gather some data, typically a small fraction of the whole network, to analyze. Thus, it is very important to examine how robust the topological features of the blogosphere are when incorporating different time durations and ways in crawling the data. This work shows that, for the two different samples of blogosphere, BlogPulse and TREC, in spite of the low overlap in their coverage and time durations of collecting, some topological features, such as the degree distributions, average shortest paths, connectivity, clustering coefficient and reciprocity show great consistency.

This work also shows that as the time duration of a crawl is extended, the features start to converge. This tells us that by obtaining some fairly comprehensive samples of the blogosphere, one can start to obtain good estimates of the topological features of the whole space.

We also examined the effects of the existence of splogs in the blogosphere, and found that splogs contribute a fair fraction of the total links volume in the blogosphere, and consequently affect the degree distributions greatly. Moreover, by looking at blogs in some large hosing sites, we find that blogs within the same hosting site are more likely to be connected than blogs in different hosting sites. However, this does not mean that there are few links outside of blog hosting sites. Rather many of the links originating at large hosting cites point to blogs with their own domain names.

For understanding the topological structure of the blogosphere, we further compared the features with some other large networks, such as the Web graph and some specific social networks. We find that they share some similarities, such as in-degree distributions, the small-world effect, and overall connectivity. However, they differ in other aspects, such as the out-degree distributions and level of clustering.

This chapter already shows that some vertices play a much more important role in the network than others. A natural question that arises from this observation is whether it is possible for us to get valuable information from a small subset of 'important vertices' in the network without knowing the entire of the network. We examine this problem in the next chapter.

CHAPTER III

Important Vertices in Networks

3.1 Introduction

In the previous chapter, we raised the question of whether some vertices play a more important role than others. In this chapter we examine whether one can create "graph synopses" using subgraphs of important vertices. To study the flow of information, to optimize engineering systems, to design efficient algorithms [22, 54, 55], and to investigate social structure and interaction, we study the statistical and graph properties of *entire* networks, including such features as degree distributions, connectivity, diameter, clustering properties, and evolution of such networks [6, 23]. For a variety of online networks, small subsets of vertices are relevant for efficient algorithms and dominate various graph and statistical properties. Frequently, these smaller subsets or *graph synopses* are easier to study and to understand. One might be interested in whether relationships among web pages can be described without crawling the whole web graph and might be inferred from a small set of vertices. We might also study the "communication" among the most influential political blogs [3] and determine whether information flows directly among them or through intermediate blogs. Despite these examples, there is little principled study of the properties, the construction, and the utilization of subsets of special vertices or edges in large

real networks. Such a study is challenging because it is hard to define precisely what is meant by a small version of the graph. Also, it is difficult to evaluate the quality of a compressed graph.

We would like a simple, principled approach to graph synopsis for a number of reasons. First, there are a number of online networks in which a synopsis of the graph is sufficient to capture the relevant information we seek. For example, rather than continuously tracking millions of blogs, one may use occasional snapshots of the blogosphere to construct a subgraph of the most "important blogs" according to a desired measure, and crawl, query, and analyze this smaller synopsis. The synopsis will allow us to capture predominant features of the much larger underlying graph, but, due to its small size, can be stored much more efficiently and even distributed and replicated amongst a number of resource-constrained computers which themselves can execute queries on the content and links.

To build a principled approach to graph synopsis, we start with the definition of predominant vertices and define a precise construction of a graph synopsis from these. Typically, the subset of vertices which capture the graph features are those which are "important." Furthermore, the importance of these vertices is highly skewed—only few of them are of great importance and the majority are less important. These vertices and subgraphs have been studied extensively in online networks [127, 27], but not with the idea of using them for graph synopses. Following much of this work, we choose four standard definitions of importance: degree, betweenness, closeness and PageRank. We demonstrate empirically for a number of representative online networks that these subsets of vertices do not depend highly on the choice of importance measure. Next, we show that it is possible to glean accurate information about the communication, relationship, and flow of information on the original graph

and among the top vertices simply from a subgraph constructed from the important vertices. Furthermore, these properties are consistent, regardless of the importance measure we use, and are appropriate for efficient algorithm design and information management.

We give a clear, precise definition of the algorithmic problem of *vertex-importance graph synopsis* in Section 3.2 and discuss the computational hardness of this problem in Section 3.4. We show in Sections 3.3 and 3.4 that most online networks are far from the worst-case graph; they exhibit features (e.g., power-law degree distribution, short average diameter, and high clustering) that allow us to efficiently compute a graph synopsis. Moreover, we tie properties of the subgraphs to measures, such as assortativity, of the original networks. Finally, in Section 3.5, we match the empirical observations to analytical results.

3.2 Preliminaries

3.2.1 Importance measures

The definitions of *importance* or *prominence* on vertices vary significantly depending on the specific network and application. Most such measures describe the topological location of the vertices. We choose four of the most commonly used measures in various applications as our objects of study: *degree*, *betweenness*, *closeness*, *PageRank*.

Let the graph $G(V, E)$ have $|V| = n$ vertices, the four importance values defined on vertices v_i are listed below:

1. **Degree** $D(v_i)$: previously defined in Chapter II, is a measure of how many vertices in G are connected to v_i directly. If G is a undirected graph, then $D(v_i)$ is the number of undirected edges incident to v_i; if G is a directed graph, then

$D(v_i)$ is the sum of indegree and outdegree of v_i, where indegree is a count of the number of directed edges to the vertex, and outdegree is the number of directed edges from that vertex to others. Degree reflects a local property of the vertices in the graph.

2. **Betweenness** $B(v_i)$: a measure of how many pairs of vertices go through v_i in order to connect through shortest paths in G:

$$B(v_i) = \sum_{j<k} g_{jk}(v_i)/g_{jk}$$

where g_{jk} is the number of shortest paths linking vertices j and k; and $g_{jk}(v_i)$ is subset of those paths that contain vertex v_i. For a directed graph G, the shortest paths are directed shortest paths. Betweenness reflects a global property of the vertices in the graph.

3. **Closeness** $C(v_i)$: a measure of the distances from all other vertices in G to vertex v_i:

$$C(v_i) = \left[\sum_{j\neq i} d(v_i, v_j)\right]^{-1}$$

where $d(v_i, v_j)$ is the distance between v_j and v_i. Intuitively, closeness means that vertices that are in the "middle" of the network are important. For a directed graph G, the closeness of a vertex could be computed in three ways: all directed paths *to* the vertex, all directed paths *from* the vertex, and all paths regardless of direction. In our work we use this third version, effectively treating the graph as undirected.

4. **PageRank**: a variant of the Eigenvector centrality measure and assigns greater importance to vertices that are themselves neighbors of important vertices [86].

3.2.2 Description of network datasets

We chose our network data sets to be representative of web and online social network data for which one might be interested in examining the properties of important vertices and their graph synopsis. We complement three empirical data sets with analysis of Erdös-Renyi (ER) random graphs, in order to discern interesting features in real world graphs from patterns that may arise by chance. For directed and undirected graphs, we measure the properties of the directed or undirected versions respectively, restricting ourselves to the largest weakly connected component.

Table 3.1: The average shortest path (ASP) and other characteristics of the largest components of the graphs.

	Erdös-Renyi	`BuddyZoo`	`TREC`	Web
Vertices	10,000	135,131	29,690	152,171
Edges	49,935	803,200	195,940	1,686,541
ASP	4.26	5.96	3.72	3.48
Directed	False	False	True	True

Erdös-Renyi random graph. An Erdös-Renyi random graph is a prototypical random graph with each pair of vertices having an equal probability p of being joined by an edge. In our model, we set the number of vertices $|V| = 10000$ and choose $p = \frac{1}{1000}$, so the average degree is $\langle d \rangle = p \times |V| = 10$.

Budyzoo dataset. The first real-world network we consider is derived from the website `buddyzoo.com`. The system, no longer active, allowed users to submit their AOL Instant Messenger (AIM) buddy lists to compare with others.

By treating each registered user as a node and their Buddy List as a series of edges to other nodes, a graph is formed. Our anonymized snapshot of the data from 2004 includes 140,181 registered users [47]. In this chapter, we keep only reciprocal ties (74.7% of the total edges), producing an undirected graph.

TREC. The second real-world graph considered is a network of blog connections,

the TREC (Text REtrieval Conference) Blog-Track 2006 dataset [73]. It is a crawl of 100,649 RSS and Atom feeds collected over 11 weeks, from December 6, 2005 to February 21, 2006. In our experiments, we removed duplicate feeds and feeds without a homepage or permalinks. We also removed over 300 Technorati tags, which appear to be blogs, but are in fact automatically generated from tagged posts, and so are not true indicators of social linking. The TREC dataset contains hyperlinks of various forms, including blogrolls, comments, trackbacks, etc. There are 198,141 blog-to-blog hyperlinks in total, and 33,385 blogs having at least one such link.

Web graph dataset. The web graph data set was collected in 1998 by Alexa[1] and has previously been analyzed as part of the "Web in a box" project at the Xerox Palo Alto Research Center [2]. Since the snapshot was collected such a long time ago, it contains *only* 50 million pages and 259,794 websites. This "small" size allows us to comprehensively analyze the web graph. We construct a directed graph where Site A has a directed edge to site B if any of the pages within A point to any page within site B.

Due to the similarity of results for the recent blog datasets and the decade old website-level data set, we expect our results to be applicable to larger, more current webcrawls.

3.3 Important vertices

In this section, we examine the graph synopsis consisting of important vertices in the network. First, we describe some properties of the entire networks. Second, we analyze the subgraphs induced by important vertices. Finally, we compare some properties of the important vertices in the subgraphs and the entire networks.

[1] `www.alexa.com`

3.3.1 Network properties and important vertices

Degree distributions. We plot the cumulative degree distributions of three real online networks in Figure 3.1. We treat the Web and `TREC` networks as directed graphs and plot the distributions of their in-degrees and out-degrees and we treat `BuddyZoo` as an undirected graph. By fitting the distributions of in-degree of Web and `TREC` with power-law distributions, we get their power-law exponents, which are 2.47 and 2.16 respectively. Moreover, we can see that the degree distribution of `BuddyZoo` has a very sharp drop off at the tail, which is observed in many social networks, e.g., co-authorship networks [79]. This places blog links, a form of social linking, somewhere between navigational/informational general linking on the Web and the reciprocal, communicative linking of a social network. The distributions of out-degree of Web and `TREC` show mild deviations from power laws, consistent with other web measurements [90] and might be due to the limitation of the data sampling [101].

Correlation of importance values of different measures. Before examining the important vertices in the networks, we look at the relationships of importance measures in different networks. Table 3.2 shows that all of the importance measures are positively correlated in all four networks. The two undirected graphs, Erdös-Renyi and BuddyZoo, have more highly correlated importance measures. Perhaps the directed edges of the other graphs add complexity to centrality measures. Furthermore, we see that, for all of the networks, *degree*, *betweenness* and *PageRank* have higher correlation than *closeness*. Thus, we see that there are various ways of defining importance in the networks and the most *central* vertices according to different centrality measures share overlap significantly.

Assortativity. The concept of *assortativity* or assortative mixing is defined as

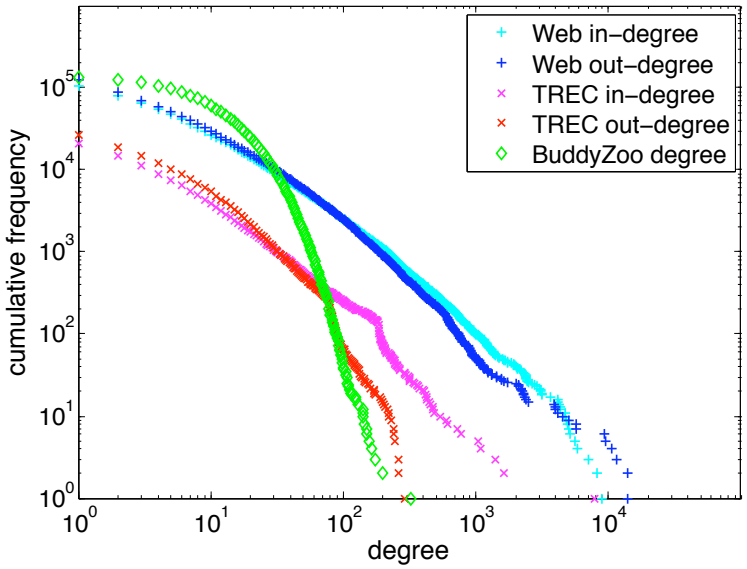
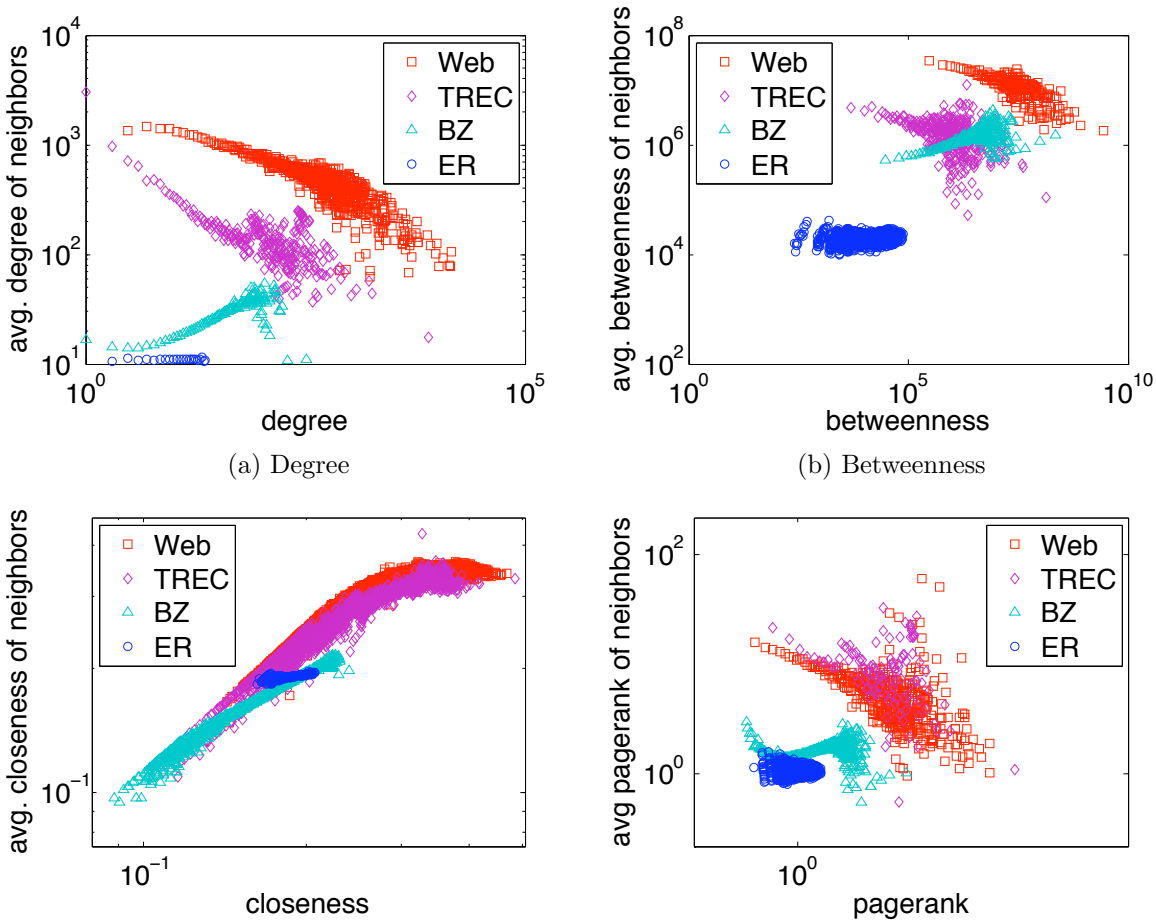


Figure 3.1: The degree distributions of online networks of `BuddyZoo` data, `TREC` blog data and Web data.

Table 3.2: Spearman correlations between importance measures of vertices. All the p-values of the correlations are < 0.0001.

Correlation	Erdös-Renyi	`BuddyZoo`	`TREC`	Web
Deg, Bet	0.9920	0.8137	0.7872	0.6178
Deg, Clo	0.9474	0.7849	0.3835	0.7869
Deg, PR	0.9952	0.9486	0.7058	0.5175
Bet, Clo	0.9673	0.7541	0.3120	0.4709
Bet, PR	0.9823	0.8439	0.7439	0.6757
Clo, PR	0.9154	0.6418	0.1086	0.3253

the preference of the vertices in a network to have edges with others that are similar. Here, we will focus on similarity with regard to centrality. We choose to measure the average value $\langle k \rangle$ of the neighbors of vertices of importance value k, i.e. $\langle k \rangle_{neigh}(k) = \sum_{k'} k' P(k'|k)$, where k is determined by each of the four different importance measures [89]. From the change of $\langle k \rangle_{\text{neigh}}(k)$ as k increases, we deduce the network's assortativity for this particular valuation. When the overall slope of $\langle k \rangle_{neigh}(k)$ is positive, the network is assortative; if the overall slope is negative, then it is disassortative. Otherwise, the network is neutral (e.g., the assortativity of degree



(a) Degree (b) Betweenness

(c) Closeness (d) PageRank

Figure 3.2: The slopes of the distributions of $\langle k \rangle_{\text{neigh}}$ show the assortativities.

of Erdös-Renyi random graphs).

In Figure 3.2, we can see that all four networks are consistently assortative with the importance measure of closeness. This confirms our intuition—the neighbors of the vertices with high closeness also have high closeness. The other three importance measures consistently show that the Erdös-Renyi random graph is a neutral graph, that `BuddyZoo`, similar to other social networks [77], is assortative, and that the Web and `TREC` blog networks are mildly disassortative. We'll see in Section 3.5 that this result does not mean that important blogs avoid linking to other important blogs. Rather, there is such a large skew in the linking behavior to important blogs, that

one would expect at random for them to already be linking to one another very frequently.

3.3.2 Important vertices in their subgraphs

In this section, we discuss important vertices and the subgraphs induced by these vertices. Such analysis helps us to discover the information hidden behind the important vertices in the real online networks, and how we can utilize them for graph synopsis. We do not fix a specific threshold for inclusion of important vertices in the subgraph, as this may vary by application. Rather, in our study what occurs as we allow the absolute number of important vertices m vary, as long as $m \ll n$, where n is the number of vertices in the original network.

Figure 3.3 shows the subgraphs induced by the four importance measures in `BuddyZoo` and the highest degree vertices in the other three networks. These subgraphs may be markedly different for different measures of importance, even within the same graph, in spite of high correlation in importance measures among vertices. They may also vary significantly between graphs, even for the same importance measure. There are several explanations of this behavior. Given the high assortativity of the closeness measure, we are unsurprised to find that individuals of high closeness are closely connected in the BuddyZoo graph. Buddyzoo also has individuals of high degree, but there were limits imposed both by AOL and individuals' own bandwidth, and so the large connected component among high degree vertices does not contain all such vertices. On the other hand, the highest degree vertices in both the blog and web datasets have such high degree that they tend to form a single connected component.

Connectivity. The first question we address is whether the connectivity of important vertices depends on other, less important, vertices or whether they are already

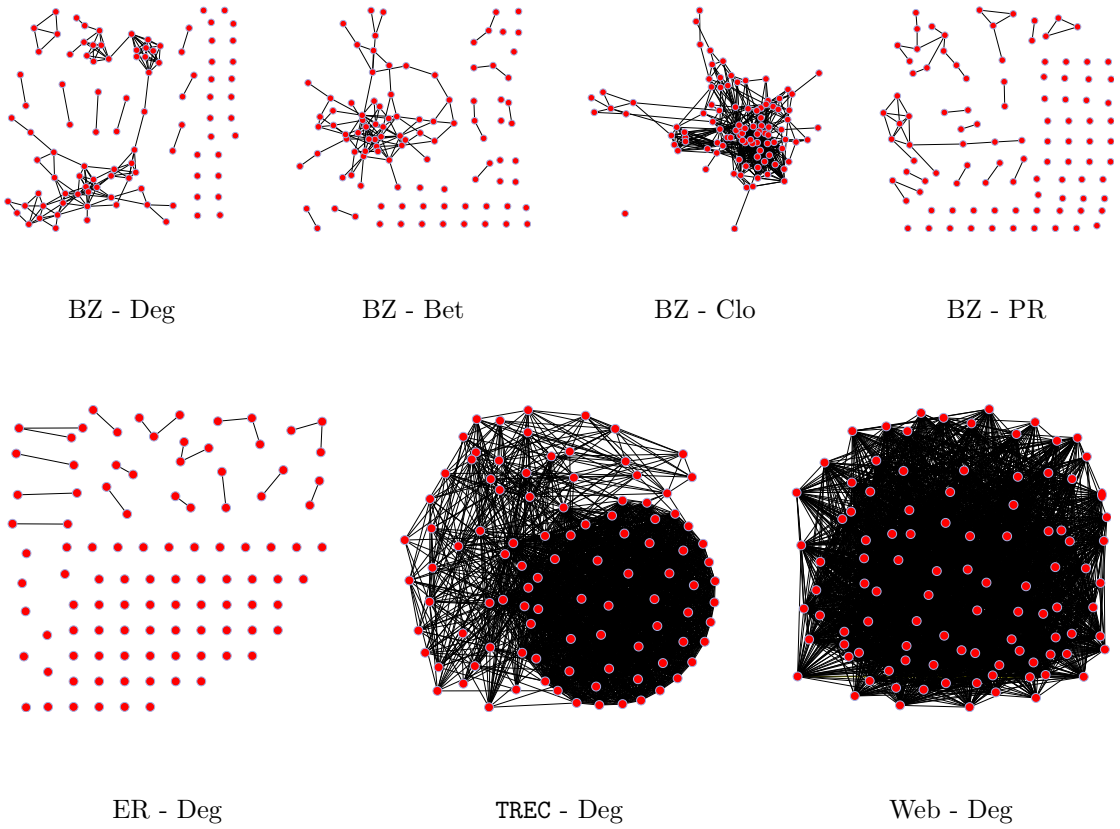


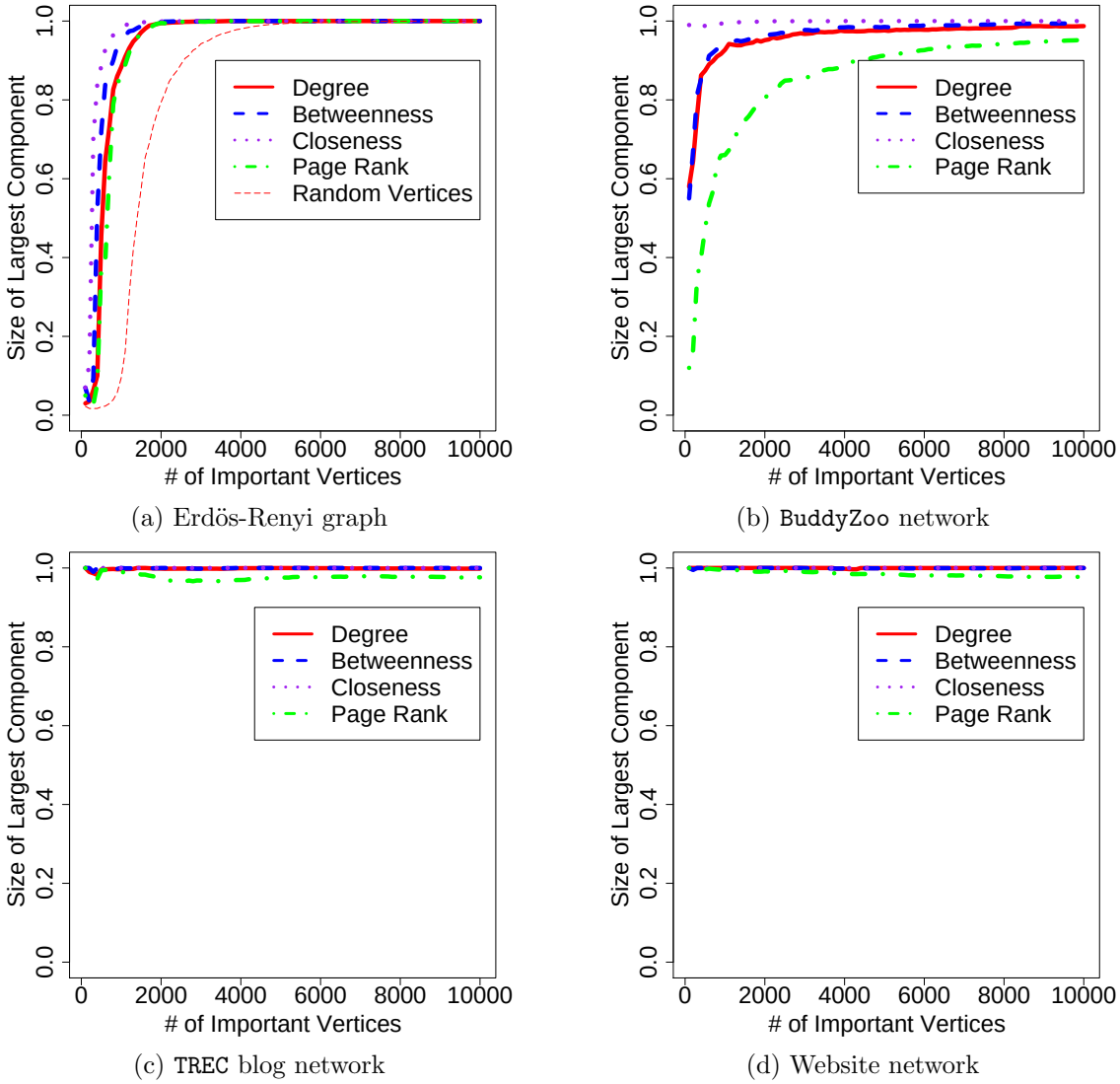
Figure 3.3: In the top row are subgraphs induced by the top 100 important vertices of `BuddyZoo` for all four importance measures, while in the bottom row are subgraphs induced by the 100 highest degree vertices in the other three networks.

well connected through one another. In the Erdös-Renyi random graph, the size of the largest connected component is given by the solution x to the equation

$$x = 1 - e^{-\langle k \rangle x}$$

where $\langle k \rangle$ is the average degree of the graph. The solution to this equation, shown as a dotted red line in Figure 3.4(a), represents the change in size of the largest connected component of the subgraphs induced by picking vertices randomly from the Erdös-Renyi random graph. When we choose vertices according to importance instead, the subgraphs have significantly better connectivities, with the largest connected component occupying 96.5% of the subgraph once the subgraph contains over 15%

of all vertices in the graph (i.e., 1,500 important vertices vs. 10,000 total vertices).



(a) Erdös-Renyi graph

(b) `BuddyZoo` network

(c) `TREC` blog network

(d) Website network

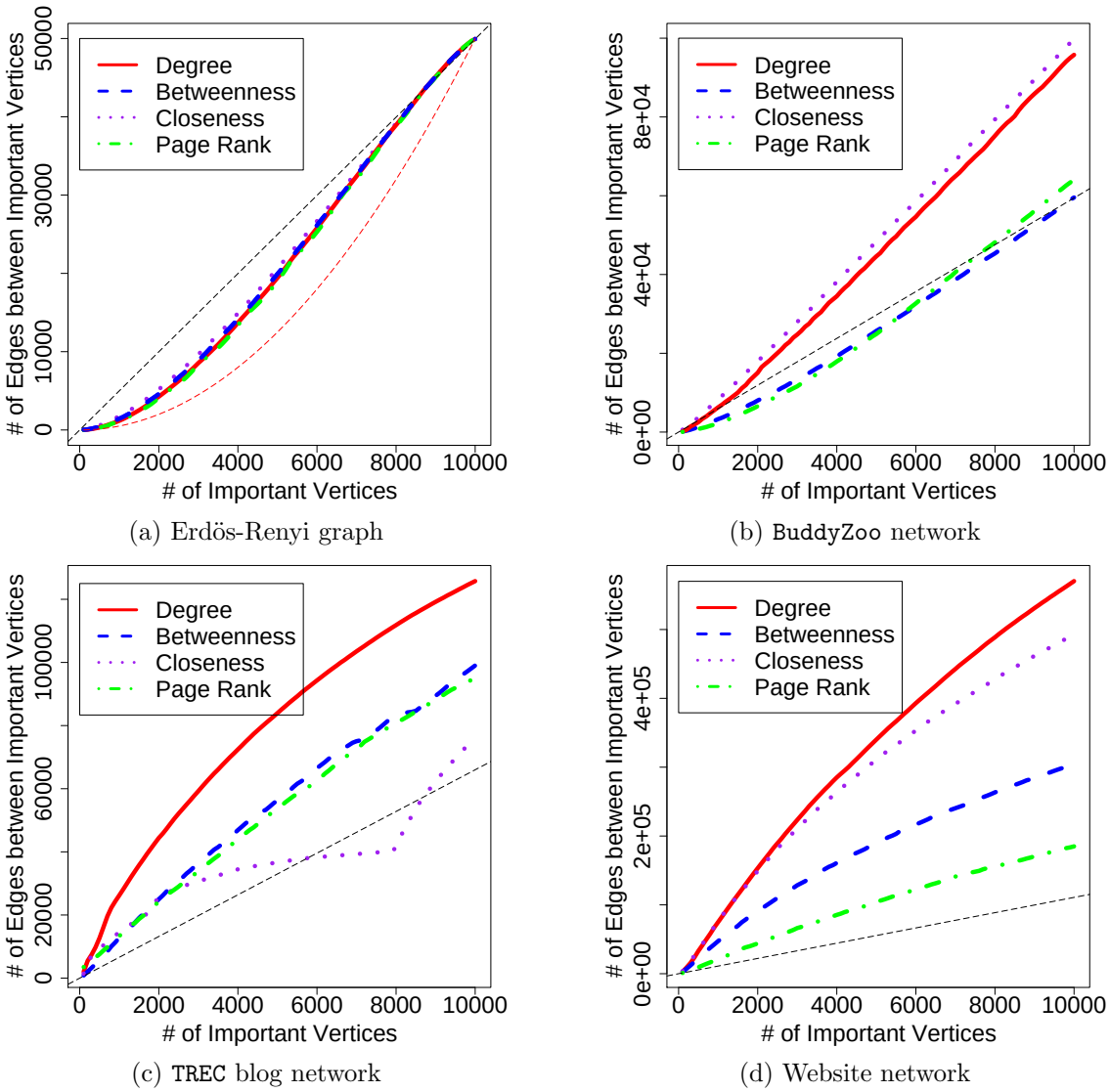
Figure 3.4: The sizes of largest connected component of the sub-networks of important vertices in Erdös-Renyi random graph and three real online networks.

Moreover, from Figure 3.4, we see that the important vertices are even more highly connected in the real networks. No matter which network and which importance measure, all of the curves of the connectivity of important vertices are almost monotonically increasing. For `BuddyZoo`, more than 95% of the important vertices of highest degree, betweenness or closeness are in the largest connected component

when they comprise just 1% percent of all vertices in the network (i.e., 1,500 important vertices vs. 135,131 total vertices). In addition, more than 95% of the 10,000 highest PageRank vertices are in the largest connected component. For both of the two directed networks, the TREC blog network and the network of websites, the most important vertices are very well connected ($> 99.5\%$) even when their numbers are very small ($< 0.05\%$ of all the vertices in the networks). Note that this very high level of connectivity is in spite of the dissortative nature of the TREC and website networks with respect to degree, betweenness and PageRank, where important vertices tend to connect to less central vertices. We can reconcile the two by observing that the important vertices are already interconnected, so the negative assortativity comes from highly connected vertices being connected to lower degree vertices simply because they already have so many connections and there is only a small percentage of vertices of similarly high degree [88].

Density. The previous observations tell us that the connectedness of important vertices is high even when we omit all other vertices in the original graph and even when they comprise a very small fraction of the entire network. Next, we examine just how dense their connections are. In Figure 3.5, we show the relationships between the number of edges incident on important vertices and the number of important vertices.

Figure 3.5 (a) shows that for an Erdös-Renyi graph, the important vertices according to all four measures have a higher average degree in the subgraph than randomly chosen vertices (red dashed line), but this average degree is lower than the average degree in the complete graph (black dashed line). The density of the graph reaches a maximum when all of the vertices in the graph are included. Moreover, from the direction of the curves, we can see that the number of edges e increases super-linearly



(a) Erdös-Renyi graph

(b) `BuddyZoo` network

(c) `TREC` blog network

(d) Website network

Figure 3.5: The growth of numbers of edges between important vertices. The slope of the black dash line in each plot is the ratio of the number of edges v.s. the number of vertices in the entire network.

with the number of important vertices n, i.e. $\Theta(n) < e < \Theta(n^2)$.

However, Figures 3.5(c) and (d) reveal the opposite behavior for networks with highly skewed degree distributions (TREC and Web). The curves of each network do not overlap as much, and the average degree of the important vertices in the subgraph is higher than the average degree in the original network. This indicates that rather than being sparser, as was the case for the Erdös-Renyi subgraphs, the subgraphs

of important vertices in real world online networks are actually *denser*. Finally, for the BuddyZoo network (Figure 3.5 (b)), which is assortative, but not power-law in degree, we see a mix of trends. Subgraphs of vertices with high betweenness and PageRank tend to be a bit sparser than the complete network, but the most important vertices according to degree and closeness are more densely connected (this is also apparent in the visualizations in Figure 3.3).

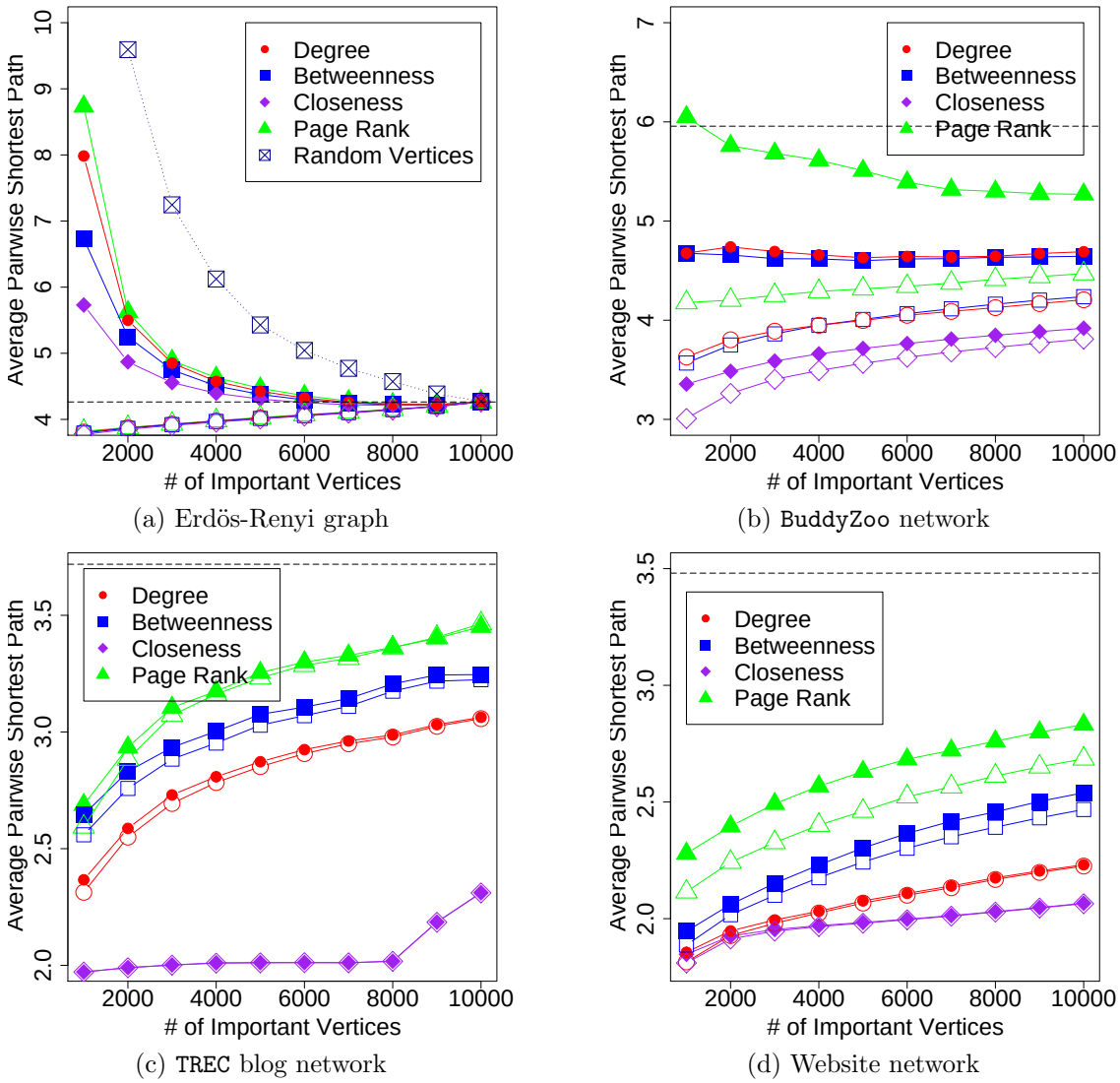
In examining these real online networks, we see that although the densities of connection among important vertices vary considerably in different networks with different importance measures, in general, they are significantly denser than random vertices in the Erdös-Renyi random graph.

3.3.3 Original vs. subgraph properties

Distance. In Section 3.3.2 we saw that even without any additional vertices from the original graph, the subgraphs of important vertices in the three online networks are already well connected. Next we examine the second property that we want to preserve for our graph synopsis problem: the average shortest paths (ASP) between reachable pairs of important vertices.

Figure 3.6 shows the comparison curves of ASPs of important vertices in their induced subgraphs and in the original networks. In the Erdös-Renyi random graph, the ASP between important vertices is on average shorter than the ASP for the entire network (the dotted baseline). But in their induced subgraphs there are significantly more hops between them on average, which indicates that important vertices in random networks are not closely connected, and their shortest paths route through non-important vertices. Nevertheless, subgraphs of important vertices in ER graphs are somewhat better connected than subgraphs of randomly selected vertices.

In contrast to the Erdös-Renyi random graph, all three real online networks con-



(a) Erdös-Renyi graph

(b) BuddyZoo network

(c) TREC blog network

(d) Website network

Figure 3.6: The ASP of: all vertices in the entire networks (black dashed line), important vertices in the the subgraphs (solid points), important vertices in the entire networks (hollow points).

sistently show that the ASPs between important vertices are much shorter than the average shortest paths of the entire network; and almost all of them are increasing as more important vertices are added in. What is more, by comparing the ASPs of important vertices in the original graphs and in the subgraphs, we see that their values are extremely close in most cases, especially for the TREC and Web data, e.g., the solid and hollow purple points (ASP of vertices of highest closeness) are almost ex-

actly overlapped. This indicates that important blogs are most efficiently connected through other important blogs.

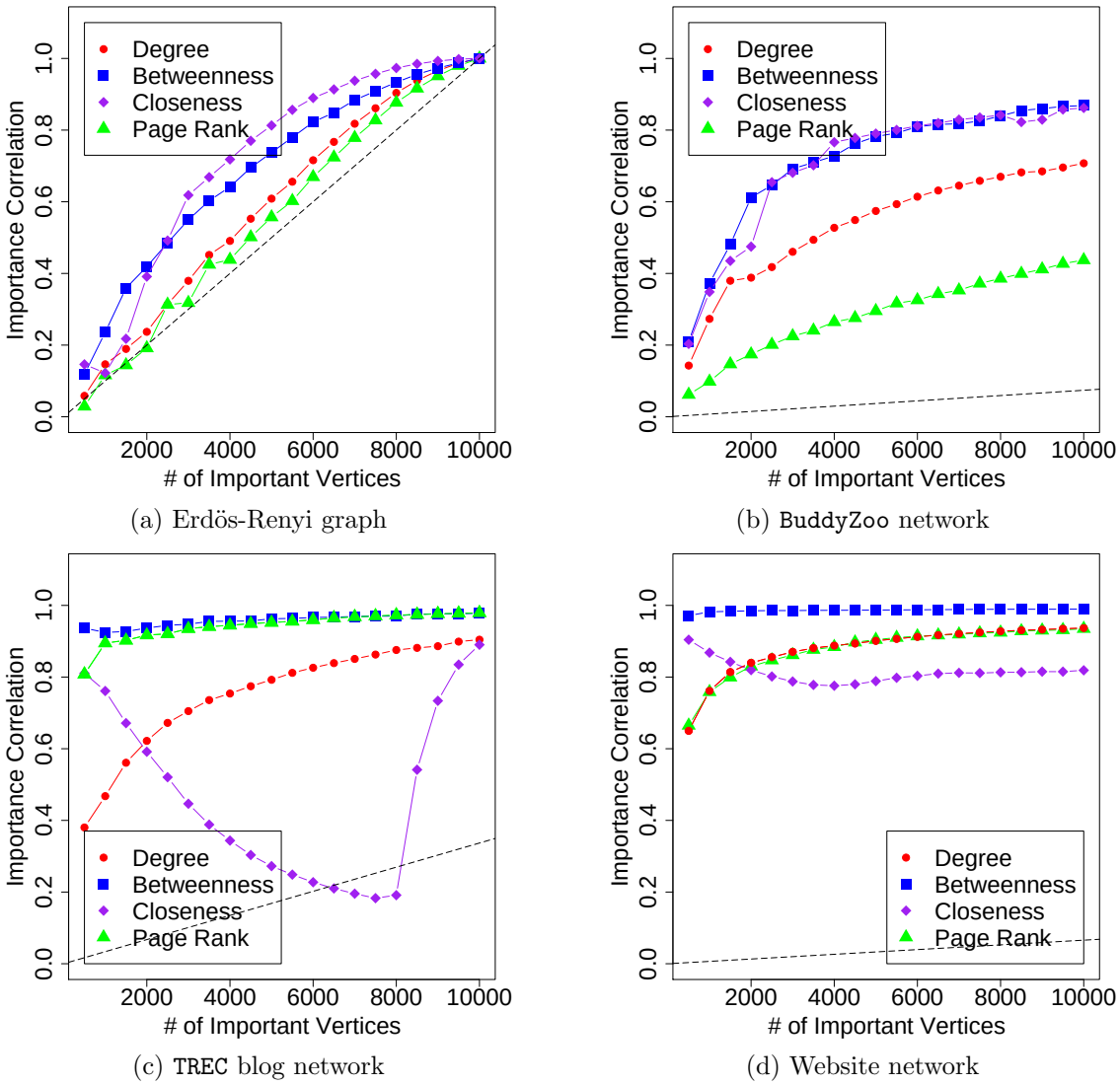
Relative importance. In addition to the connectedness of important vertices, we are also interested in their relative ranking: if we only keep the important vertices and the edges among them, how would the vertices rank in the new subgraph with the same importance measure? In order to answer this question, we generate subgraphs of different sizes for all networks. We then compute the importance of the vertices in the subgraphs according to the same importance measure used to select them. Finally, we compute the Pearson correlation of the importance values of those vertices in the original graph and in the subgraph.

Figure 3.7 shows that the correlations are all much higher for the real-world online networks than the Erdös-Renyi random graph, and that this is especially true for the Web and TREC data. The abnormality of closeness in TREC may be due to the blog aggregators and splogs. There are one to two thousand blogs whose only incoming link is from centrally positioned (in the network) blog aggregators. This boosts the closeness score of the unimportant blogs, creating the abnormality in Figure 3.7.

The high correlations of the online networks tell us that the ranking of importance in the subgraphs of important vertices is highly consistent with their ranking in the original graphs. This suggests that, e.g., it may not be necessary to crawl all blogs to get an accurate ranking of the most important blogs. Rather, the links among the top blogs themselves may already provide fairly close approximate rankings.

3.3.4 Summary

After studying the important vertices and their induced subgraphs, we can make two overall observations about the four networks: (i) Different importance measures yield subgraphs of varying density and topology as is evident in Figure 3.3. (ii)



(a) Erdös-Renyi graph

(b) `BuddyZoo` network

(c) `TREC` blog network

(d) Website network

Figure 3.7: Pearson correlations of importance values of vertices in subgraphs and original graphs. The black dashed lines are the base lines starting from 0 when the number of vertices is 0; and ending at 1 when all the vertices in the networks are included.

However, in spite of these differences, "important vertices" in the online networks have some properties that agree with each other, which are essential for the graph synopsis we are looking at: they connect to each other more directly than average; their distances to each other are closer than between random pairs of vertices; and their relative ranks are positively correlated to their importance ranks in the original networks. Thus, we know that in the real online networks, in contrast to random

graph model, the subgraphs induced by the important vertices tend to preserve information about the relationships among important vertices, and we can use the subgraphs to study the properties of important vertices in the original graphs.

3.4 Compression with guarantees

While retaining only the important vertices may be sufficient to capture most of the relationships among them in real-world networks, in general we have no guarantee that these induced subgraphs preserve any properties at all (whether of the important vertices or of the original graph). We cannot even guarantee the most basic property of connectivity of the important vertices. In this section, we rigorously define the graph compression problem, analyze the computational complexity of two heuristic algorithms, and discuss the trade-offs of these approaches.

3.4.1 Hardness of compression with guarantees

We define the Basic Graph Compression Problem as follows: In a connected unweighted graph $G(V, E)$, every vertex is assigned an importance value. Taking the original graph $G(V, E)$ and the set of vertices S with largest importance values as inputs, find the minimal set of additional vertices ν, which form a connected subgraph $G'(V', E')$, where $V' = S + \nu$ and $V' \subseteq V$, $E' \subseteq E$.

We recall the Network Steiner Tree Problem which is NP-complete [38]. A heuristic method, the *Minimal Spanning Tree* algorithm gives solutions to this problem with approximation ratio 2 [40]. One can show that Basic Graph Compression and the Network Steiner Tree Problem are polynomial-time reducible to one another. Thus, Basic Graph Compression is an NP-complete problem.

3.4.2 Heuristic algorithms

There are, however, several heuristic algorithms that guarantee the preservation of some properties of the important vertices in the original graph. We detail the KeepOne and the KeepAll algorithms [39] next, and note the similar web projection method [66].

KeepOne. Let K_1 be the set of important vertices, the goal is to find the minimal set K_2 such that there is a tree induced by $K_1 \cup K_2$. The approximation algorithm is first to build a minimum spanning tree on the complete graph on K_1 where an edge (u, v) has weight equal to the length of a shortest path from u to v. The set K_2 consists of any additional vertices along any "path" edge in the minimum spanning tree. The result is the graph induced by the vertices $K_1 \cup K_2$.

The KeepOne algorithm guarantees the connectivity of the compressed graph, has the same set of additional vertices as the projection method in [66], and only introduces more edges, which means it may have better diameter preservation than the projection method.

Unfortunately, retaining only connectivity may provide a distorted view of the original graph. We see in Figure 3.8 an example of a graph on n vertices in which the distance of the original vertices a and b is 3 but in the compressed graph built by KeepOne, their distance is $n-3$. The ratio of the distances is $\frac{n-3}{3}$ which we can make arbitrarily large by increasing the number of vertices n. That is, KeepOne retains connectivity but may drastically distort the distance between some pairs of important vertices. To ameliorate this problem, one can use the KeepAll algorithm [39] which keeps vertices that lie along a shortest path between any two vertices in K_1.

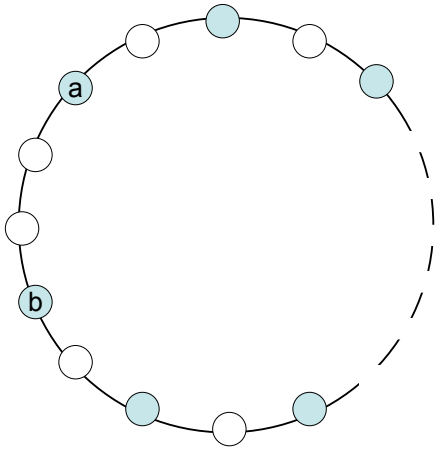


Figure 3.8: The distance of important vertices a and b in the original graph is 3 and $n-3$ in the compressed graph obtained by KeepOne. The ratio of distances can be made arbitrarily large as $\lim_{n\to\infty} \frac{n-3}{3} = \infty$.

3.4.3 Empirical evaluation and trade-offs

While Figure 3.8 shows that the worst case distance preservation of KeepOne may be arbitrarily bad, real-world networks are far from the worst case. Furthermore, the KeepOne and KeepAll algorithms illustrate that there are some tradeoffs we may make in compressing real-world graphs—we can maintain distances at the cost of keeping a few additional vertices. To explore these empirical tradeoffs, we apply both the KeepOne and KeepAll algorithms to three networks. Table 3.3 shows these results. Since the results with the Web data are very similar to `TREC`, we do not list them here for conciseness. From the table, we can see that if we insist on preserving the pairwise shortest paths of all important vertices, we must include many more additional vertices (thus increasing the size of our synopsis). Furthermore, we must do so even though the average pairwise shortest paths in the subgraph of just the important vertices is already close to that of the original graph. Note that we increase the size of the synopsis by fewer than 100 additional vertices when we preserve connectivity (with KeepOne), but we need over 3000 additional

Table 3.3: Comparison of the properties of subgraphs generated by different methods with important vertices in Erdös-Renyi random graph, BuddyZoo and TREC. Sub-Importance Measure100 is the subgraph induced by top 100 important vertices only; KO- is the subgraph generated by KeepOne; KA- is the subgraph generated by KeepAll. LC is the fraction of important vertices in the large component of the subgraph. Avg PSP is the average pairwise shortest path length in the subgraph.

Subgraph	Add vts	LC	Avg PSP	Subgraph	Add vts	LC	Avg PSP	Subgraph	Add vts	LC	Avg PSP
Erdös-Renyi				BuddyZoo				TREC			
Sub-Deg100	0	NA	NA	Sub-Deg100	0	0.58	NA	Sub-Deg100	0	1	1.636
KO-Deg100	80	1	14.526	KO-Deg100	33	1	9.440	KO-Deg100	0	1	1.636
KA-Deg100	3222	1	3.649	KA-Deg100	2199	1	3.233	KA-Deg100	34	1	1.609
Sub-Bet100	0	NA	NA	Sub-Bet100	0	0.55	NA	Sub-Bet100	0	1	2.085
KO-Bet100	68	1	15.497	KO-Bet100	35	1	16.087	KO-Bet100	0	1	2.085
KA-Bet100	3185	1	3.633	KA-Bet100	2376	1	3.171	KA-Bet100	216	1	1.994
Sub-Clo100	0	NA	NA	Sub-Clo100	0	0.99	2.599	Sub-Clo100	0	1	1.716
KO-Clo100	62	1	11.474	KO-Clo100	1	1	2.624	KO-Clo100	0	1	1.716
KA-Clo100	3000	1	3.604	KA-Clo100	531	1	2.324	KA-Clo100	0	1	1.716
Sub-PR100	0	NA	NA	Sub-PR100	0	0.12	NA	Sub-PR100	0	1	1.298
KO-PR100	87	1	15.404	KO-PR100	75	1	11.517	KO-PR100	0	1	1.298
KA-PR100	3338	1	3.672	KA-PR100	3978	1	3.880	KA-PR100	36	1	1.294

vertices when we also insist on preserving distances. In short, while the problem of preserving connectivity in graph compression is NP-complete, heuristic algorithms such as KeepOne can preserve connectivity with a lower cost, while preserving the distances demands quite more. In this sense, we can also see that the short pairwise shortest paths of important vertices in their subgraphs and their original graphs is a special and important property of the online networks we study.

3.5 Analytical discussions

In this section we present the expected density of subgraphs of random graphs with varying degree distributions, in order to contrast these expected values with the empirically observed measurements. We limit ourselves to vertex degree as the sole importance measure and assume that the graphs are random aside from the degree distribution, which we specify. We then obtain the density of the subgraph by deriving the probability that an edge in the original graph lies between two vertices in the subgraph.

First, we find the degree k_i of the least important vertex among the set of top i

most important vertices. We do so by calculating the expected number of vertices of degree at least k_i in a network with $n = |V|$ vertices. Furthermore, we assume that the expected number is actually equal to i so that

$$i = n \cdot P(k_i).$$

Because we are given the complementary cumulative distribution function $P(k)$ explicitly for Erdös-Renyi and power law random graphs, we can solve the previous equation for k_i and, after doing so, we find the probability that an edge is incident to a single important vertex, $e \to V_i$, given by

$$\mathbb{P}(e \to V_i) = \frac{1}{|E|} \int_{k_i}^{n} k \cdot p(k)\, dk$$

where $p(k)$ is the pdf of the degree distribution. Using independence of the edges, we find that the number of edges within the subgraph of important vertices is simply

$$|E_i| = |E| \cdot \mathbb{P}(e \to V_i)^2.$$

3.5.1 Erdös-Renyi graphs

In an Erdös-Renyi random graph, the degrees are distributed according a Poisson distribution where the probability of a vertex having degree larger than the mean decreases exponentially. As a result, even when selecting the highest degree nodes, their degree will be within an order of magnitude of the average degree $z = \langle k \rangle$ of the network.

In Figure 3.9, we show the number of edges in the subgraph of an Erdös-Renyi graph, using the normal distribution with mean z and standard deviation $\sigma = \sqrt{z/n(1 - z/n)}$, is

$$i = \frac{1}{2}\left(1 + \operatorname{erf}\Big(\frac{k_i - z}{\sigma\sqrt{2}}\Big)\right).$$

We see that when the number of important vertices is small, the degree within

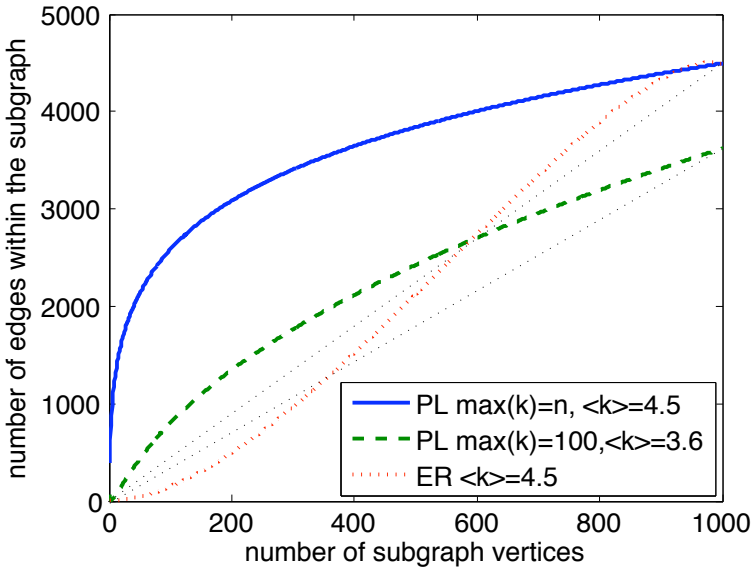


Figure 3.9: The number of edges between important vertices, where importance is measured by degree, in three networks: 1) power law network with $\alpha = 2.2$, $n = 1000$, 2) Erdös-Renyi graph with the same average degree, and 3) power-law graph with the same exponent but a cutoff at $k = 100$. Two dotted lines show what the number of edges would be if the average degree in the subgraph were equal to the average degree in the original network.

the subgraph is lower than the degree of the original graph. Using well known properties of Erdös-Renyi graphs, we expect that when the average subgraph degree is 1, a giant component will emerge in the subgraph, and further, when the average degree is $\log(n)$, the subgraph will be path connected. This is consistent with the set of connectivity and density measurements on simulated Erdös-Renyi graphs in Section 3.3.2.

3.5.2 Power law graphs

We expect different behavior in power law graphs, where high degree vertices are so well connected that they will naturally connect, not only to a large portion of the network, but also to one another as well. For example, in a power-law graph

with exponent α and no cutoff on the degree[2], one vertex on average is expected to have degree $N^{1/(\alpha-1)}$ [81]. For $\alpha = 2$, this means that one expects one node to be connected to majority of the other nodes.

In selecting high degree nodes in a power law graph, we are selecting nodes that are likely to be connected to each other by virtue of the fact that so many edges are incident on them. The number of vertices with degree k_i or greater is given by

$$i = n\ P(k \geq k_i) = \frac{n}{k_i^{\alpha-1}}$$

Solving for k_i, we have that the degree of the i^{th} most important vertex is $k_i = (\frac{n}{i})^{\frac{1}{\alpha-1}}$. Next, we want to find out what proportion of the edges are incident on the i most important vertices. For this we have

$$P_e(i) = P(e \in e_i) = \frac{\int_{k_i}^{n} kp(k)dk}{\int_{1}^{n} kp(k)dk} \tag{3.1}$$

$$= \frac{k_i^{2-\alpha} - n^{2-\alpha}}{1 - n^{2-\alpha}} = \frac{(\frac{n}{i})^{\frac{2-\alpha}{\alpha-1}} - n^{2-\alpha}}{1 - n^{2-\alpha}} \tag{3.2}$$

Figure 3.9 shows that the average degree in the subgraphs of important vertices is actually *higher* than in the original graph. We repeat the analysis using a degree distribution cutoff $\max(k)$ that is lower than the total number of nodes n. This cutoff not only disallows very high degree vertices, but also lowers the average degree in the original subgraph. When the cutoff is introduced, the subgraph still maintains a higher average degree than the original graph, but the difference is less pronounced.

Note the similarity with Figure 3.5, showing the number of edges in the subgraph for the TREC and Web data sets, both of which are power law in nature (although directed). In both the analytical and empirical subgraphs, the average degree is higher than it is for the entire graph. We should mention that for exponents $\alpha \sim 2$ and very small i, Equation 3.2 would yield a higher average degree than there are

[2]a cutoff may be imposed such that $P(k) \sim k^{-\alpha}$ for $k < \max(k)$ and 0 otherwise

important vertices to connect to. This is in fact a known property of random power law graphs, where simply fixing the degree of a vertex and allowing it to satisfy this degree by forming edges at random would create a non-vanishing frequency of multiple edges between highly connected vertices. If one disallows multiple edges, the networks become mildly disassortative, consistent with our empirical observations.

3.6 Related work

In this section, we examine the graph sampling problem and the rich-club phenomenon. Both of them have some similarities with our problem: the former also studies how to get "good" subgraphs given large massive networks; and the later focuses on the set of "important vertices". However, they are still different from our problem in various aspects. In graph sampling, one aims to devise a sampling method, e.g., random vertex or edge selection, snowball sampling, the sketching-based sampling [69] etc., in order to be able to infer the properties of the original graph from the much smaller sampled graph [62, 67]. In contrast, our work constructs subgraphs of predetermined important vertices, not for the purpose of deducing properties of the original graph, but in order to infer the underlying relationships amongst the important vertices themselves.

In the "rich-club phenomenon", vertices with high degree tend to connect together tightly, which is true for many social and other types of real networks [127, 27]. While previous work on the rich-club phenomenon has aimed to determine whether the number of edges between high degree vertices based purely on degree is higher than what one would expect at random, our study extends to other centrality measures, and describes essential properties of the subgraphs themselves, such as connectivity, shortest paths, and preserving rank orderings of importance. A related analysis of

highly interconnected sub-structures in networks is that of k-cores, subgraphs of vertices where each vertex has at least k connections within the subgraph [34]. An interesting direction for future work would be to repeat our analysis of the properties of the subgraph and original graph, using k-core membership as the importance measure for vertex selection.

3.7 Conclusion

In this chapter, we propose a new approach to analyzing and studying large online networks, *vertex-importance graph synopsis*. Given a set of important vertices, we extract a much smaller subgraph from the original network, containing those important vertices. We attempt to place this process on a rigorous footing and show that even simple versions of the graph compression problem are hard (but that there are reasonable heuristic algorithms). Unlike previous methods which evaluated the fidelity of the "graph abstract", this approach utilizes the subsets of important vertices and edges and the information they could provide in large networks. We argue that they can make information accesss and management more efficient in real applications. These observations suggest future work in using graph synopses for information retrieval and information flow detection.

From our empirical analysis of three real online networks, we find a number of interesting properties. The important vertices are much more closely and densely connected to each other. They also have significantly shorter pairwise paths, which do not heavily depend on the rest of vertices in the networks, (i.e., their pairwise shortest paths in the subgraphs induced by themselves are close to those in the original graphs). Finally, their relative ranks are almost all highly correlated to their ranks in the original networks. Although our experiments show that the properties of

vertices of different importance measures in different networks do vary in some ways, the observations stated above are consistent no matter the type of networks (either social or technological), and regardless of the importance measure we choose. Thus, we may use vertex-importance graph synopses as small but accurate representatives of the important vertices in the larger graph (and, sometimes, of the larger graph itself). Furthermore, the real online networks are relatively easy to compress while preserving important graph properties (they do not exhibit the worst-case behavior of our theoretical analysis).

In addition to empirical studies, we use analytical discussions to show how these properties of important vertices in online networks differ from random graph models. What is more, we also use heuristic algorithms to measure the complexities and trade-offs of requiring some properties of the real networks to be guaranteed in the compressed graphs.

In this and the previous chapters, we have analyzed some structural features associated with vertices in information sharing networks. While it is important to know whether two vertices are connected, directly or indirectly, in information sharing networks the strength of the ties is also a key determinant in whether information will flow. In the next chapter, we are going to see some features about the edges in the information sharing networks.

CHAPTER IV

Strong Ties in Networks

4.1 Introduction

The strength of weak ties is the concept that individuals tend to be more successful in acquiring information about job opportunities by contacting their weak ties: the individuals that they do not see often [44]. The rationale behind this idea is that close friends tend to have similar information because they share similar interests, profession, or geographical location. Weak ties on the other hand are between individuals who don't have much in common, including other contacts, and the information they have access to will tend be different. In this sense, it has been assumed that weak ties play a key role in transmitting information rapidly and widely in social networks. Here, for the first time, we challenge this assumption through a structural analysis of networks where the weak ties are removed.

Our motivation for considering networks without weak ties is that there are many situations where one may wish to use only trusted contacts to gather or disseminate information. For instance, one may be interested in assembling a team or otherwise gathering information that is distributed in different parts of a social network using only strong ties. In the case of the Madrid terrorist bombings on March 11th, 2003, the individuals behind the attack were able to procure knowledge about making

explosive devices, hashish to trade for explosive materials, and the explosive material itself using their strong ties. Had they used weak ties which would have been less reliable, their plot may have been exposed and their intentions thwarted. Sinister plots are not the only example of a planning activity that can benefit from using strong ties to maintain confidentiality. Scientists may wish to forge collaborations requiring diverse expertise [46], and in doing so they may wish to keep a competitive edge by not broadcasting their ideas over weak ties. Similar situations may arise in the formation of business alliances, where companies seek to complement their strengths through mergers, acquisitions, cross licensing of intellectual property, or joint ventures, but do not wish to leak their next steps to competitors.

There are also processes which describe the contagion of new ideas and practices in which the credibility of information or the willingness to adopt an innovation requires independent confirmation from multiple sources. Unlike a 'simple' biological contagious agent carrying a disease, which can be transferred through a single contact between two individuals, ideas and opinions ('complex' agents) may need to be heard from multiple contacts before being adopted [26]. The presence of closed triads in the social network, consisting of three individuals who all know one another, enhances the probability that complex contagion can spread on the network. As two neighboring contacts are infected, they have a greater probability to infect their shared contacts who will now be hearing about the news or product through from multiple sources. Complex contagion may apply to processes ranging from teenagers adopting a new brand of jeans to farmers starting to plant a new type of corn [98]. In these scenarios, the decisive event may not be hearing about an innovation, but observing enough people participating to be convinced that the innovation should be adopted [111].

Given that processes such as sharing of sensitive information and adoption of

certain innovations may only occur via strong ties, we study the connectivity and small-world property of social networks consisting entirely of strong ties. In different contexts the strength of a tie may have different definitions and measures, such as frequency or length of contact. For simplicity, in this chapter we consider only the presence of closed triads as evidence of "strong ties". This is based on the assumption that good friends or close professional contacts will know at least some people in common. Throughout this chapter, "weak ties" are taken to be those that are not part of any closed triad and "strong ties" are the ones that share at least one other contact in common.

Social networks tend to have a much higher probability of closed triads than the equivalent random networks [122, 81]. An intuitive reason is given by structural balance theory [25] which states that ties tend to be transitive: if a node is connected to two other nodes (is a member of two diads), those two nodes are much more likely on average to be connected than two randomly chosen nodes. Recently, it has also been shown that many real world networks, including social networks, contain overlapping k-cliques [87]. Within a k-clique, each of the k nodes is connected to each of the other $k-1$ nodes, forming a densely knit community containing $\binom{k}{3}$ closed triads. Two cliques were considered overlapping if they shared $k-1$ nodes, and the question was posed whether these overlapping cliques themselves form a network containing a fraction of the network (the network percolates). In contrast, in this work, we are interested not in the overlap of cliques, but the strength of ties between individuals. A message can be passed between two communities, even if they share only one individual in common, as long as that individual has strong ties within both communities. Therefore our condition of transitive edges between two information sharing nodes is less restrictive than the requirement that the cliques

themselves contain a very high degree of overlap.

Our results are as follows. Given the potential importance of closed triads both in assembling varied expertise and in the diffusion of innovation, we first determine how they are linked together in observed social networks. We find that transitive ties are prevalent in social networks and removing non-transitive ties from these social networks shrinks the giant component, but does not break it up. This result indicates that social networks are composed of overlapping communities, with each community providing strong ties, and the overlap providing a way to traverse the network using strong ties. Besides measuring the properties of real world networks with weak ties removed, and we also model random networks consisting entirely of closed triads. This allows us to quantify the impact this local structural requirement has on the global properties of a network, such as the connectivity of the network and the small-world properties.

Previous work [78, 24, 48, 112, 11, 12] has modeled networks with varying degrees of clustering. However, our very simple model is the first to explicitly address how requiring *all* ties to be transitive affects network properties. To this end, we model a random graph constructed entirely of closed triads and compare its properties to that of an Erdös-Renyi graph with the same number of nodes and edges. We derive both theoretically and numerically the result that the giant connected component occurs at the same average connectivity (average degree $\langle k \rangle = 1$), but that it does not grow so quickly in the triad graph as the average connectivity increases further. Numerical simulations reveal that the average shortest path is quite similar in both networks. Essentially, requiring transitive closure allows fewer nodes to be connected (since 1/3 of the links must be redundant rather than reaching out to connect additional nodes). However, the resulting connected component will have an average shortest

path that scales logarithmically with the size of the graph, just as it would in an Erdös-Renyi graph.

The remainder of this chapter is organized as follows. In Section 4.2 we present an empirical analysis showing that social networks (online friendship networks in this case) are not dependent on weak ties to stay connected through a short number of hops. In Section 4.3 we compliment the empirical analysis with a random graph model that preserves the connectivity and small world properties of an Erdös-Renyi graph while satisfying the condition that each tie be transitive. This model demonstrates that one need not sacrifice much in the way of connectivity within the network in order to satisfy the requirement of transitivity.

4.2 Online social networks without weak ties

In order to study the connectedness of social networks without weak ties, we analyzed two data sets. The first, and smaller data set is the social network of the Club Nexus online community at Stanford in 2001 [1]. Much like many later online social networking services, it allowed individuals to sign up and list their friends on the site. The 'buddy' lists were aggregated into a single social network of reciprocated links. Within a few months of its introduction, Club Nexus attracted over 2,000 undergraduates and graduates, together comprising more than 10 percent of the total student population. The Club Nexus network is only a biased subset of the complete student social network because students had free choice of how many friends to list. Nevertheless, the data does provide a proxy of the true social network, from which one can derive interesting properties. For example, triangles are quite prevalent in this network, with a clustering coefficient of 0.17, which is 40 times greater than what it would be for an equivalent Erdös-Renyi random graph. The

average distance between any two individuals is just 4 hops.

Adamic et al. [1] found that edges with high betweenness, where betweenness reflects the number of shortest paths that traverse the edge, tended to connect people with less similar profiles. These profiles included information about the student's year, field of study, personality, hobbies and other interests. The observation that ties of high betweenness lie between dissimilar individuals supports the hypothesis that weak ties bridge different communities. Edges with high betweenness also tend not to be part of closed triads, because each edge in the triad provides a possible alternate path. In fact, a recently-devised clustering algorithm relies on identifying communities by removing edges that participate in fewest closed triads and longer loops [91]. It is therefore a concern that removing non-transitive ties from a network would tend to break it apart into disconnected communities. This would mean that diverse expertise may not be reachable and new innovations may not flow throughout the network.

In the case of the Club Nexus network, we can dismiss the concern, because the network is robust with respect to the removal of weak links, which account for 19% of all links. Rather than breaking up into many disconnected communities, the network sheds some nodes and shrinks modestly. Most obviously, the 239 leaf nodes cannot be part of triangles because they link to just one other node. They each become a disconnected component with the removal of weak ties, which is justified in this context because they are peripheral actors. Table 4.1 shows the distribution in size of the connected components for the original network and the network with weak links removed.

Note that both networks have a giant component containing the majority of the nodes. The removal of weak ties does not separate communities of large size—the

component size	Club Nexus	Club Nexus without weak ties
2246	1	0
1763	0	1
6	0	1
5	1	1
4	1	2
3	2	4
2	8	0
1	227	710

Table 4.1: Distribution of connected components in online communities.

largest one is composed of just 6 nodes. The removal of weak ties does cause a slight increase in the the average shortest path between reachable pairs. Although the fraction of reachable pairs drops from 72% to 51%, the average shortest path increases from 3.9 hops to 4.1.

The next network we consider is the network of AOL Instant Messenger (AIM) links submitted to the website `buddyzoo.com`. The system uses Buddy Lists to show users which buddies they have in common with their friends, to visualize their Buddy List, to compute shortest paths between screennames, and to show each user's prestige based on the PageRank [86] measure applied to the network. Our anonymized snapshot of the data is from 2004 and includes 140,181 users who submitted their buddy lists to the BuddyZoo service, as well as 7,518,816 users who did not explicitly register with BuddyZoo but were found on the registered users' Buddy Lists. This is therefore a rather large social network. It was previously studied to determine whether direct links can be concealed in the network, for example to manipulate an online reputation mechanism [47]. In the context on BuddyZoo, this would mean that two people would remove each other from their Buddy Lists in an attempt to hide their connection. But unless they share no other 'buddies' in common, they would still be linked as 'friends of friends' and arguably would have a more difficult time denying acquaintance. Nine percent of the users have only a single connection,

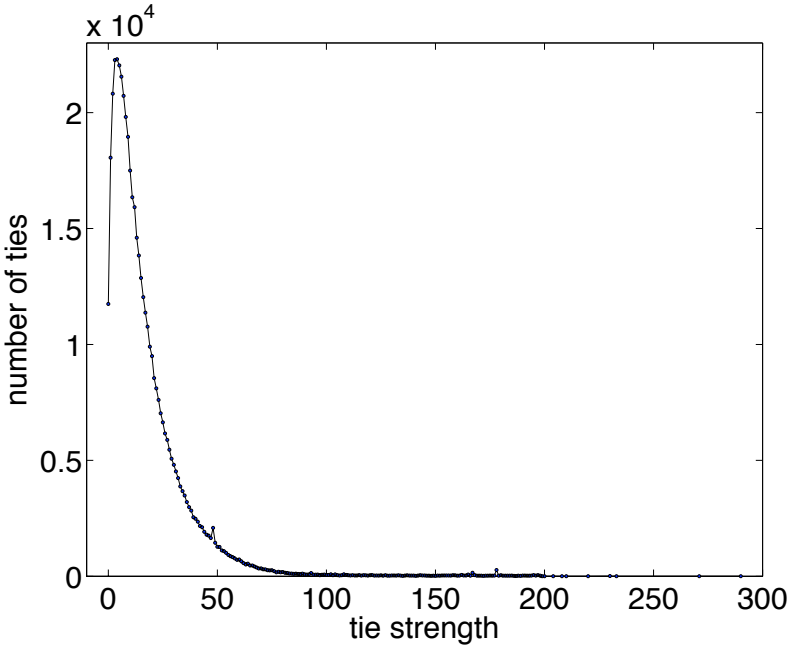


Figure 4.1: The distribution of the strength of ties, measured as the number of triads each tie participates in.

and would disconnect themselves from the network if they were to remove it. Of the remaining pairs of users, only 19% could remove their direct link and be at least distance 3 from each other, while all others would remain friends of friends. This is equivalent to asking what percentage of the edges are parts of triangles, which is the question we are currently interested in.

In order to determine the presence of strong ties, we consider only users who explicitly registered with BuddyZoo, but we allow an edge to be considered transitive if it is part of a closed triad that includes an unregistered user. This is because we know that two people share a contact, even if that contact did not register. We exclude 9 shared contacts that have indegree greater than 1000, because those could be AIM bots (automated response programs). Even disregarding the 23 contacts that have an indegree greater than 300 (corresponding to the average size of a typical person's offline network [74]), does not affect the results significantly. We do not include

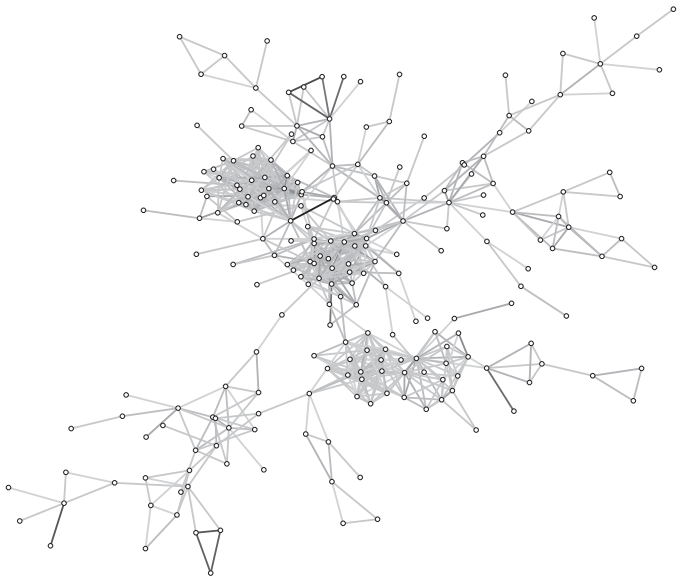


Figure 4.2: The largest component of the reduction of the BuddyZoo network where each tie participates in at least 47 triads. The triads themselves are not all shown — only the ties that share a threshold number of them.

unregistered contacts in the network itself because their Buddy List information is incomplete. The degree distribution is highly skewed and there are many isolates in the network. On average, each user is connected via a reciprocated tie to 6.83 other registered BuddyZoo users. We require a tie to be reciprocated, since it is possible for one AIM user to add someone to their buddy list without that person adding them in turn.

As in the case of the Club Nexus social network, we find that removing weak ties does not have a dramatic effect on the BuddyZoo network. Although several communities containing a couple of dozen nodes do split off, the giant component shrinks modestly, from occupying 88.9% of the graph to occupying 87.5% of it. The average shortest path increases by a fraction of a hop from 7.1 to 7.3. Usually any lengthening in the path decreases the probability of a successful transmission if the probability that the message is transferred at each step is less than 1 [121]. How-

component size	BuddyZoo	BuddyZoo without weak ties
124672	1	0
122066	0	1
21-40	0	1
11-20	11	14
10	4	6
9	5	5
8	7	9
7	7	10
6	15	16
5	37	36
4	64	73
3	126	168
2	591	685
1	7279	9413

Table 4.2: Distribution of connected components in the BuddyZoo AOL instant messenger community. A tie is considered weak if two users who list each other on their buddy lists do not list a third person in common.

ever, we do not observe considerable lengthening of the average shortest path until we impose a higher threshold on tie strength. In order to consider more restrictive requirements on tie strength, we vary the strength threshold as follows: rather considering any tie in a single closed triad to be strong, we require that it be part of at least j closed triads. Figure 4.1 shows the distribution of tie strengths, where the mean number of shared ties is 17.4 and the median is 13. Figure 4.2 shows the largest component of nodes where each tie participates in at least 47 triads. There are several dense cliques, but the largest component is quite small—only 233 nodes. To investigate how rapidly the giant component shrinks and how much the average shortest distance changes, we consider reduced networks where only ties of above threshold strength, measured by the number of triads the tie participates in, are kept. Figure 4.3 shows the giant component size and average shortest path between all connected pairs as the threshold is increased from zero to 35 triads. We observe that the giant component shrinks gradually, indicating that a substantial portion of the network is spanned by ties of moderate strength. This would indicate that

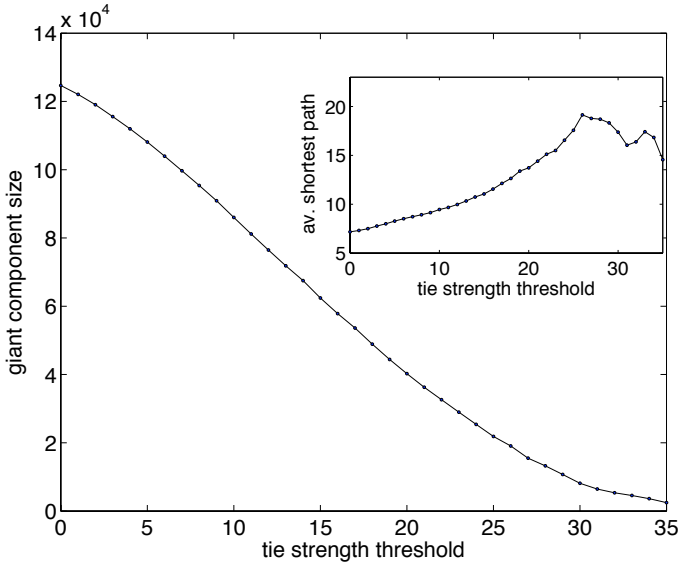


Figure 4.3: The size of the giant component as only ties of a minimum strength (measured in the number of triads it is a part of) are kept in the network. The inset shows the growth of the average shortest path between connected pairs.

the network is composed of overlapping communities rather than separate communities that are bridged by weak ties. What is more, removing weak ties does not separate large communities from one another. Rather, a few smaller communities and many isolates are spun off as the tie strength threshold is increased. Removing weak ties has an additional cost beyond isolating some individual nodes and smaller communities—it increases the average shortest path between reachable pairs. So even though the giant component is shrinking, we are removing the shortcuts that span it. The average shortest path more than doubles as we increase the threshold from 1 to 25.

The strong tie robustness of the Club Nexus and BuddyZoo networks is encouraging, especially in comparison to what one might expect in a Watts-Strogatz (WS) type small world model [122] or an Erdös-Renyi graph. In the WS model, the network is constructed from a lattice where each node is connected to k neighbors on each side. For $k > 1$, this means that each node participates in local closed triads. In

the model, a fraction p of the links are rewired with one endpoint placed randomly among the nodes. It is the presence of these random links that gives the WS model a shortest path that scales logarithmically with the size of the graph. Such a link is unlikely to be part of triangle however, since the probability of any two nodes linking randomly is proportional to $1/N$ in such a graph. Therefore, removing weak links in a WS model removes the shortcuts, leaving an average shortest path that scales linearly with the size of the graph. Assuming that nodes close together on the lattice share similar information, one would need to make many hops in order to find novel information. In Section 4.3.3, we will show that the occurrence of strong ties in an Erdös-Renyi graph is unlikely unless the average degree increases with the number of nodes in the network. Therefore, removing all edges that are not part of a triangle will isolate most of the nodes in random graphs where the average degree is constant or nearly constant with respect to the number of nodes.

4.3 Random graphs composed of strong ties

Given the empirical results of the previous section, where we see a very high prevalence of transitive ties and a robustness of the network with respect to removal of weak ties, we seek to answer the basic question of what the cost is of requiring all ties to be transitive. We measure this cost in terms of the connectivity and average shortest path of a network where every edge between two nodes is part of at least one closed triad and compare to the equivalent Erdös-Renyi graph, where no transitivity constraint is imposed.

To this end, we construct the simplest random graph composed entirely of triangles, and we model this kind of graph by assigning links simultaneously among any three randomly chosen nodes in the graph. Strictly speaking, for a graph with

$|V| = N$ nodes, there are $\binom{N}{3}$ possible combinations of nodes that can form a triangle. Each triangle forms with probability b, so that on average we randomly choose $M = b \times \binom{N}{3}$ triplets of nodes and link them with three edges. Our method of constructing transitive graphs is similar to a particular instance of the Newman [78] model for constructing highly clustered graphs. In the Newman clustered network model, one takes a bipartite network of individuals and groups. One then constructs a one-mode projection of the random graph by adding, with a given probability p, edges directly between individuals who belong to the same group. However, unlike [78], in our model the probability for nodes to connect to each other in the same group is 1, and the number of members in each group is constant at 3.

4.3.1 Degree distribution

We consider the degree distribution of the graph starting from the distribution of a node belonging to k closed triads.

For each node u, there is a total of $R = \binom{N-1}{2}$ possible triangles which have u as one of the vertices. And, for each triple of vertices, the probability of being selected to have links in the graph is b. Let r_m be the probability for a node belong to m chosen triples. Then

$$r_m = \binom{R}{m} b^m (1-b)^{R-m}. \tag{4.1}$$

On the other hand, we will now show that it is unlikely that our fixed node u is part of two triangles with an edge in common. Our node u has degree k if, for some m, node u is in m chosen triples on a total of k distinct nodes aside from u. It is straightforward to show that $k/2 \leq m \leq \binom{k}{2}$. In fact, for even $k \ll N$, most of the probability is in the case $m = k/2$. For even k, the probability that u has degree

k is the probability that u is in exactly $k/2$ chosen triples, adjusted for collisions of edges. Collisions affect the probability of degree k in two ways—u may be in exactly $m = k/2$ triples but a collision reduces the contribution to the probability of degree k, or u may be in $m > k/2$ chosen triples but collisions increase the contribution to the probability that the degree is k.

We consider a node u belonging to m triples involving j neighbors and consider the probability of a collision occurring. Conditioned on u falling in exactly m chosen triples, all sets of m triples are equally likely. There are $\binom{R}{m} = \Theta\left(\frac{N^{2m}}{2^m m!}\right)$ possible sets of m triples. Next, we want to count the number of sets of m triples involving exactly j neighbors of u, for $j \leq 2m$. We can pick the j neighbors as a set in $\binom{N-1}{j}$ ways, but then we need to assign roles to the j neighbors based on collision multiplicity. For example, suppose 4 triples among five neighbors A, B, C, D, E of u might be $\{u, A, B\}, \{u, A, C\}, \{u, A, D\}, \{u, B, E\}$. We can choose A, B, C, D, E as a set; pick an element for the role of A (that appears three times) in 5 ways; given that, pick an element for the role of B in 4 ways; then E in 3 ways, and the remaining elements take the interchangeable roles of C and D, for a total of $5 \cdot 4 \cdot 3 \leq 5!$ orderings).

For us, a crude bound for the orderings of roles will suffice. There are at most $2m - j$ collisions counting multiplicities, and so at most $2m - j$ neighbors of u that can be in more than one triple—and can play a non-trivial role. There are at most $2m - j$ roles. So the number of ways to assign non-trivial roles is at most $(2m - j)^{2m-j}$. So the number of sets of m triples involving exactly j neighbors of u is at most $\binom{N-1}{j}(2m - j)^{2m-j}$. Thus the ratio of these to the number of sets of m

disjoint triples is

$$\begin{aligned}\frac{\binom{N-1}{j}(2m-j)^{2m-j}}{\binom{R}{m}} &\leq O\left(\frac{N^j(2m-j)^{2m-j}2^m m!}{j!N^{2m}}\right)\\ &\leq O\left(\frac{((2m-j)/N)^{2m-j}2^m m!}{j!}\right).\end{aligned}$$

We are intereseted in the case $2m-j \geq 1$. If m and j are constants, then we can ignore $2^m m!/j!$, and we get

$$\begin{aligned}\frac{\binom{N-1}{j}(2m-j)^{2m-j}}{\binom{R}{m}} &\leq O\left(\frac{((2m-j)/N)^{2m-j}2^m m!}{j!}\right)\\ &\leq O\left(1/N\right).\end{aligned}$$

By choosing the appropriately small probability b of choosing a triple, we may assume that m and j are much smaller than N. But we cannot necessarily assume m and j are constants; for example, we may have $m!$ comparable to N. We now consider the case where j or m grows (slowly) with N, and where N is sufficiently large. If $m \leq j$, then $2^m m!/j! \leq \binom{j}{m}^{-1} \leq 1$. It follows that

$$\begin{aligned}\frac{\binom{N-1}{j}(2m-j)^{2m-j}}{\binom{R}{m}} &\leq O\left(\frac{((2m-j)/N)^{2m-j}2^m m!}{j!}\right)\\ &\leq O\left(((2m-j)/N)^{2m-j}\right)\\ &\leq O(N^{-1}).\end{aligned}$$

On the other hand, if $m > j$, then $2m-j > m$, so

$$\begin{aligned}\frac{\binom{N-1}{j}(2m-j)^{2m-j}}{\binom{R}{m}} &\leq O\left(\frac{((2m-j)/N)^{2m-j}2^m m!}{j!}\right)\\ &\leq O\left(((2m-j)/N)^{2m-j}(2m)^m\right)\\ &\leq O\left((2m(2m-j)/N)^{2m-j}\right).\end{aligned}$$

If $2m-j=1$, this is $O(2m/N) \leq N^{-1+o(1)}$. If $2m-j>1$, then, since we may

assume that $2m \ll \sqrt{N}$, we have

$$\begin{aligned}\frac{\binom{N-1}{j}(2m-j)^{2m-j}}{\binom{R}{m}} &\leq O\left((2m(2m-j)/N)^{2m-j}\right) \\ &\leq O\left(((2m-j)/\sqrt{N})^{2m-j}\right) \\ &\leq O\left(((2m-j)^2/N)^{(2m-j)/2}\right).\end{aligned}$$

This is $O\left((2m-j)^2/N\right) \leq N^{-1+o(1)}$. Thus we have obtained bounds on the probability that two triangles incident on a node share an edge.

Given that the effect of collisions is small, we get the probability of u having degree k is

$$p_k = \begin{cases} \binom{R}{\frac{k}{2}} b^{\frac{k}{2}}(1-b)^{R-\frac{k}{2}} \pm N^{-1+o(1)} & \text{if } k \text{ is even} \\ N^{-1+o(1)}, & \text{if } k \text{ is odd} \end{cases} \tag{4.2}$$

After ignoring the additive amount $\pm N^{-1+o(1)}$, the corresponding generating function is given by

$$G_0(z) = \sum_{k=0}^{R} \binom{R}{k} b^k (1-b)^{R-k} z^{2k} = [bz^2 + 1 - b]^R \tag{4.3}$$

The average degree $\langle k \rangle$ is then given by:

$$\langle k \rangle = G_0'(1) = b(N-1)(N-2) \tag{4.4}$$

And thus, we have the relationship between average degree $\langle k \rangle$ and the probability of any three nodes being connected by a triangle b:

$$b = \frac{\langle k \rangle}{(N-1)(N-2)} \tag{4.5}$$

When $\langle k \rangle = O(1)$, $b = O\left(\frac{1}{N^2}\right)$.

4.3.2 Accidental triangles and the clustering coefficient

We should notice that in our model, the expected number of triangles in the network is not exactly $b \times \binom{N}{3}$. There is the possibility of forming an "accidental" triangle, which can occur when the pairs of nodes a and b, b and c, and a and c are linked, but the triangle a, b, c was not among the $b \times \binom{N}{3}$ initially chosen triangles. The probability b' of this occurring is the probability $1 - b$ that no triangle was intentionally formed between the a, b, and c: $1 - b$ times the probability that each of the three edges does occur in a triangle other than a, b, c.

$$b' = (1 - b)[1 - (1 - b)^{(N-3)}]^3 \tag{4.6}$$

In this way, we know that the total expected number of triangles in this graph is $a \times \binom{N}{3}$, where $a = b + b'$.

Thus, the ratio between the actual number of triangles in the graph and the designed number of triangles is:

$$\Delta = \frac{a}{b} = 1 + \frac{(1 - b)[1 - (1 - b)^{(N-3)}]^3}{b} \tag{4.7}$$

However, b' is very small compared with b, when the average degree of a node in the graph is a constant independent of the growth of the total number of nodes N. Since we have shown that $b = O(\frac{1}{N^2})$, then it is not hard to see that the ratio of the probability for any three nodes to be part of an accidental triangle and the probability for them to be a triangle that is constructed by randomly choosing groups is:

$$\frac{b'}{b} = \frac{(1 - b)[1 - (1 - b)^{(N-3)}]^3}{b} = O\left(\frac{1}{N}\right) \tag{4.8}$$

Thus, we can see that when N is large, and the average degree $\langle k \rangle$ is independent of N, then the chance of forming an accidental triangle is quite small compared to the triangles randomly drawn in constructing the model. Figure 4.4 shows the relation between b' and average degree $\langle k \rangle$.

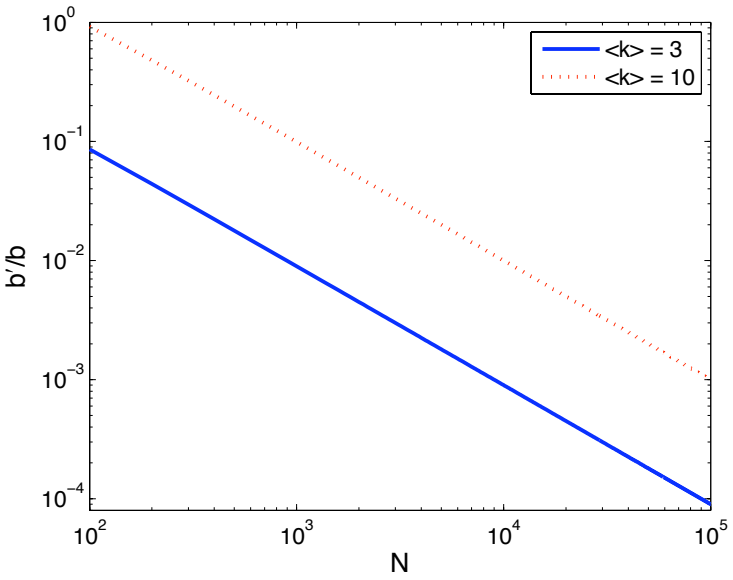
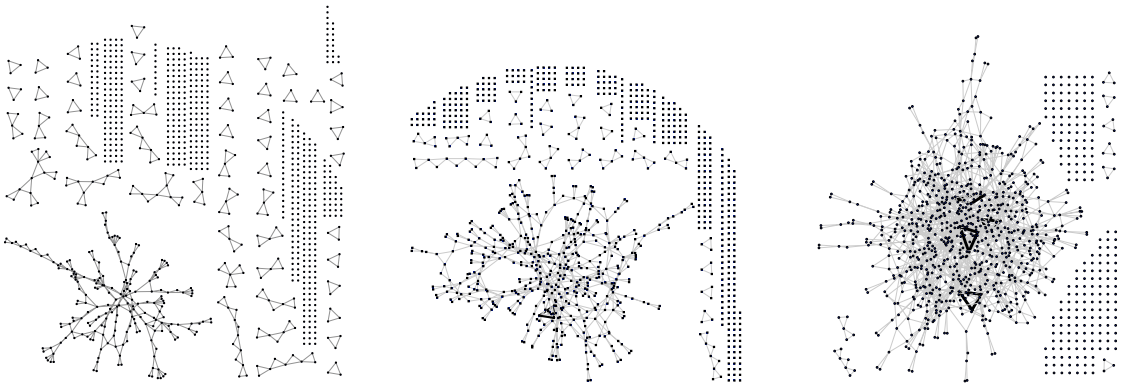


Figure 4.4: The ratio of the number of accidentally formed triangles to the number randomly chosen by the model. For fixed average degree and increasing number of nodes, the ratio of accidentally formed triangles drops as $1/N$.

In Figure 4.5 we show three instances of a randomly generated graph of triangles. Each graph has $1,000$ nodes, but we form different numbers of triangles. Even though a giant component exists for each graph, it is only once the number of triangles equals the number of nodes that we observe a few random triangles forming. Therefore the formation of accidental triangles does not have a substantial effect on the derivations below.

The clustering coefficient C is a measure of the prevalence of closed triads in a network [122, 81]. The expectation of the total number of connected triples of nodes (open and closed triads) in the graph is $N_{triple} = N \times \sum_k \binom{k}{2} p_k$, and the number of



(a) $N = 1000$, $M = 200$ (b) $N = 1000$, $M = 300$ (c) $N = 1000$, $M = 500$

Figure 4.5: Examples of triangle graphs with 1000 nodes with varying numbers of triangles M. Accidental triangles are marked with bold lines.

closed triads is $N_\Delta \approx b \times N\binom{N}{3}$ since the number of accidental triangles is small. Thus the clustering coefficient is:

$$\begin{aligned} C &= \frac{3N_\Delta}{N_{triple}} \\ &\approx \frac{3b\binom{N}{3}}{N \times \sum_k \binom{k}{2} p_k} \\ &= \frac{1}{\langle k \rangle + 1} \\ &= O(1) \end{aligned}$$

We can see that when N is large, the clustering coefficient of our graph is:

$$C = O(1) \tag{4.9}$$

which is significantly larger than the $O(N^{-1})$ clustering coefficient in an Erdös-Renyi Random graph. For many types of real world networks, it has been shown that $C = O(1)$ [81], so it is of interest to see how removing weak ties in real networks changes the clustering coefficients.

4.3.3 Phase transition and the giant component

For the derivation of the phase transition and size of giant component, we loosely follow the generating function methods for clustered graphs in [81]. The phase transition is also known as the percolation threshold—the average degree at which a finite fraction of the network is connected, forming a giant component. In Part A, we have given r_m, the probability for a node belong to m triangles. Thus, averaging over all individuals and triangles, we have the mean number of triangles a node belongs to: $\mu = \sum_m m r_m$.

The probability of having two edges within the triangle is 1, and the probability of having any other number is 0. Therefore, the generating function of the number of edges for each node within a triangle is

$$h(z) = z^2 \tag{4.10}$$

Furthermore, for a node A in the graph, the total number of other nodes in the whole graph that it is connected to by virtue of belonging to triangles is generated by:

$$G_0(z) = \sum_{m=0}^{\infty} r_m (h(z))^m \tag{4.11}$$

where r_m is the probability for a node to belong to m groups as we defined before. This is also the generating function of the distribution of the number of nodes one step away from node A.

The generating function of the distribution of the number of nodes two steps away from A is $G_0(G_1(z))$, where $G_1(z)$ is the generating function for the distribution of the number of neighbors of a node arrived at by following an edge (excluding the edge that was used to arrive at the node):

$$G_1(z) = \mu^{-1} \sum_{m=0}^{\infty} m r_m (h(z))^{m-1} \tag{4.12}$$

The necessary and sufficient condition for a giant component to exist, is when, averaging over all the nodes in the graph, the number of nodes two steps away exceeds the number of nodes one step away [82], which can be expressed as:

$$[\partial_z (G_0(G_1(z)) - G_0(z))]_{z=1} > 0 \tag{4.13}$$

Thus, we get the condition for the existence of a giant component in this graph:

$$\begin{aligned} ((\mu^{-1} \sum_{m=0}^{\infty} m(m-1) r_m z^{m-2}) \cdot h'(z))|_{z=1} &> 1 \\ 2\mu^{-1} \sum_{m=0}^{\infty} m(m-1) r_m &> 1 \\ \frac{R(R-1)b}{Rb} &> \frac{1}{2} \end{aligned}$$

After simplifying the above equation, the condition is:

$$b > \frac{1}{N^2 - 3N} \tag{4.14}$$

Since we will compare this graph with an Erdös-Renyi random graph with the same average degree $\langle k \rangle$, we express the condition for the existence of giant component in terms of the average degree given by Equation 4.4:

$$\langle k \rangle > 1 + \frac{2}{N^2 - 3N} \tag{4.15}$$

As $N \to \infty$, the condition is $\langle k \rangle > 1$. An interesting point is that this is exactly where the phase transition occurs in an Erdös-Renyi graph. Therefore, the

requirement that all edges be transitive does not delay the appearance of the giant component. It does however have a tempering effect on the rate of growth of the giant component as we will see below.

When a giant component exists in the graph and the probability for a node to whom A is connected to not belong to it is s, the size of the giant component is given by:

$$S = 1 - G_0(s_0) \tag{4.16}$$

$$= 1 - \sum_{m=0}^{\infty} r_m (s_0^2)^m \tag{4.17}$$

$$= 1 - (bs_0^2 + 1 - b)^R \tag{4.18}$$

where s_0 is the solution of the function:

$$s = G_1(s) \tag{4.19}$$

$$= \mu^{-1} \sum_{m=0}^{\infty} m r_m (s^2)^{m-1} \tag{4.20}$$

$$= (bs^2 + 1 - b)^{R-1} \tag{4.21}$$

As we have assumed $S > 0$, we know that s must be some value larger than 0 and smaller than 1, and thus $s = 1$ is a trivial solution of the function.

We compare the solution s_0 to numerical simulations of networks of random triangles. Each network contains $N = 10,000$ nodes, and we select M random triangles to connect from the N nodes. For each value of M we generate 50 random networks and average the size of the giant component. The results, shown in Figure 4.6 show excellent agreement between the analytical prediction and the numerical simulation. For comparison, we show both the numerical prediction and analytical result for the size of the giant component in an Erdös-Renyi random graph with the same number of nodes and edges. The size of the giant component in the Erdös-Renyi graph is

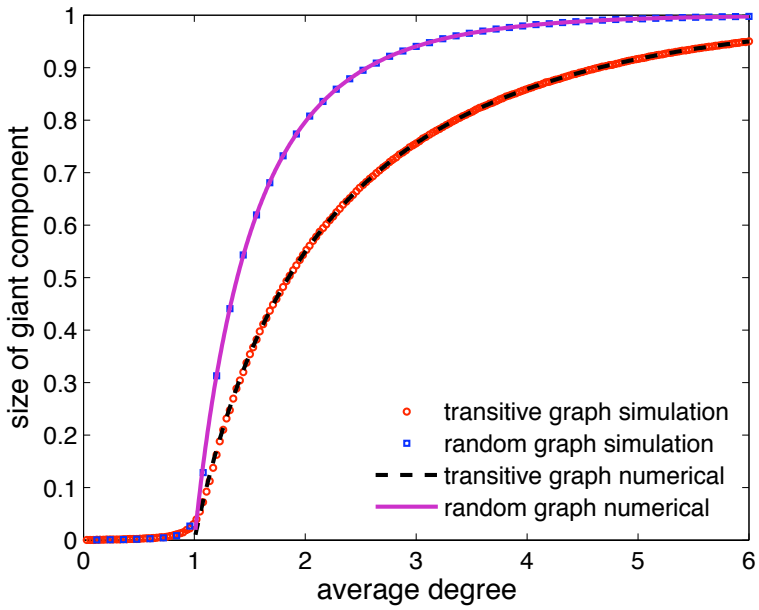


Figure 4.6: Comparison of numerical simulations with analytical solutions for the fraction of the network occupied by the giant component of a 10,000 node triangle graph and the corresponding Erdös-Renyi graph

given by the solution s to the equation $s = 1 - \exp(-\langle k \rangle s)$. From the figure, we can see that as average degree grows, the phase transitions of the transitive graph and the random graph occur at the same time, while the size of giant component of the Erdös-Renyi graph grows more quickly as we increase the average degree. An intuitive explanation is that in an Erdös-Renyi graph one need not expend a 'closure' edge to close a triad. Rather, that edge can be used to connect a disconnected node or small component to the giant component.

The fact that the phase transition occurs at the same average degree for both the Erdös-Renyi and transitive network shows that the requirement of transitivity does not result in a need for increased average connectivity in order for the giant component to form. Note that the phase transition in our model, where all edges are the result of the addition of triangles, is quite different from what it is in a graph that would result from taking a simple Erdös-Renyi graph and removing all edges

that do not fall within a triangle. In the Erdös-Renyi graph with non-transitive edge removal the percolation threshold occurs at a degree that scales as $N^{\frac{1}{3}}$.

This condition for the giant component in an Erdös-Renyi graph with weak ties removed can be derived as follows. A giant component of strong tries forms when, after arriving at an arbitrary triangle T, the expected value of the number other adjacent triangles that one could"move to" is equal to 1. The probability that there is a triangle T' adjacent to T that is not the triangle from which we reached T is given by $2\binom{N-5}{2}p^3$. There are $\binom{N-5}{2}$ choices for the vertices in T' not shared with T, and two choices of the vertex shared by T and T' (excluding the vertex of T that is shared with the triangle we arrived from). The expression $p = \langle k \rangle / N$ is the probability that any two vertices in an Erdös-Renyi graph share an edge. Thus when N is large, the average degree at the phase transition is $\langle k \rangle = N^{1/3}$. In several real world networks the average degree was found to vary as N^{β} where $0 \leq \beta \leq 0.3$ [68]. But in a random network, this density falls short of the $N^{1/3}$ necessary to make the accidental occurrence of closed triads (and therefore strong ties) high enough for the network to percolate.

If one further requires that the triangles overlap not just in one node but in two, as in the percolation of k-cliques [30], the phase transition occurs at a critical average degree that grows as $N^{\frac{k-2}{k-1}}$, with $k = 3$. This means that the average degree has to grow in linear proportion to N in order for a giant component to form. Together, these two results show that the Erdös-Renyi random graph typically does not contain sufficiently numerous strong ties to percolate. But as we have shown in Section 4.2, real world social networks do contain many strong ties that percolate. This can be intuitively explained by the observation that new social ties typically form in the context of geographical and sociocultural settings [121]. In these contexts it is natural

that the ties tend to form closed triads rather than being added independently, as they are in Erdös-Renyi random graphs.

4.4 Average shortest paths of networks of strong ties

Exact results for the average shortest path are difficult to derive even for a random graph. We therefore used numerical simulations to measure the average shortest path between all reachable nodes as we increase the size of the network. We selected a value of the average node degree where the giant component existed, but did not take up all of the graph. At our chosen value, $M = 0.5N$, there are twice as many triangles as nodes. This constant proportion of triangles to nodes means that b, the probability of any triple of nodes being connected, falls as $1/N^2$.

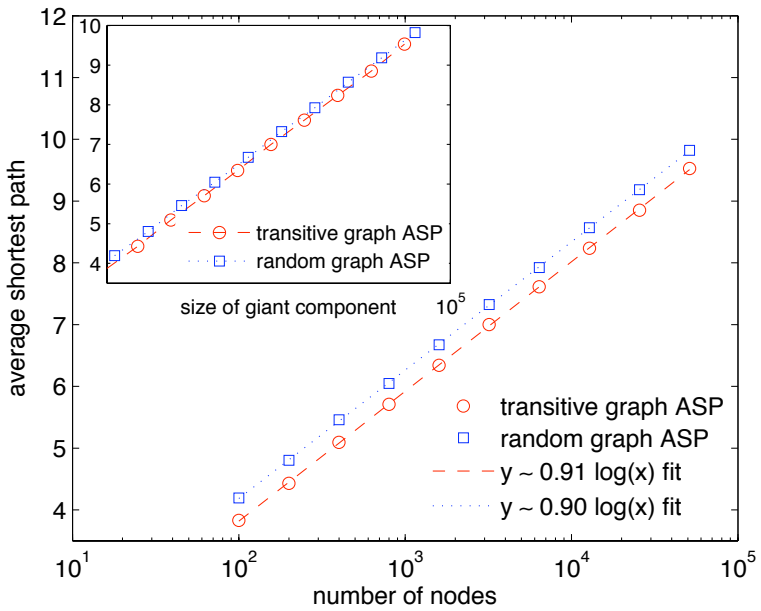


Figure 4.7: Numerical comparison of the average shortest path in triangle graphs and Erdös-Renyi graphs with the same number of nodes and edges. The inset shows the average shortest path as a function of the size of the giant component rather than the total number of nodes.

At $M = 0.5N$, the giant component occupies 76% of the nodes, while in the equivalent random graph it takes up 94% of the nodes. This makes it difficult

to directly compare the two networks, since the average shortest path is measured between reachable pairs, and the Erdös-Renyi graph has more of them. Figure 4.7 shows that the average shortest path is actually shorter in the triangle graph. This may be explained by the fact that there are fewer nodes in the giant component but a greater density of links. Once we consider the average shortest path relative to the size of the giant component, the curves become nearly identical for both networks. This shows that the requirement of triadic closure does not negatively impact the average shortest path for reachable pairs, but those pairs are fewer in number.

4.5 Conclusions

In this chapter we study the connectivity of strong ties in networks, where strong ties are defined as belonging to closed triads. We find that two real world social networks are robust with respect to removal of weak links, in the sense that there remains a giant component that is smaller but still occupies a majority of the graph. We also find empirically that the removal of weak links lengthens the average shortest path modestly. In comparison, the removal of weak links in an WS small world network or an Erdös-Renyi graph would isolate the vast majority of nodes. It is the high clustering of social networks that allows them to transmit or gather information via strong ties.

Subsequent to the original publication of this work, Kossinets, Watts and Kleinberg examined information backbones in communication networks [57]. They found that the information backbone is a sparse graph with a concentration of both strong and weak ties. This finding sheds further light on the relationship between tie strength, connectivity and information diffusion in social networks.

In the work of [85], from large-scale networks based on phone calls it is observed that there is a coupling between interaction strengths and the network's local structure. Their analysis also shows the removal of strong ties has little impact on the network's overall integrity, while this is not the case for weak ties.

We also pose a basic question, which is the cost paid for the requirement of transitive ties in terms of the size of the giant component and the length of the average shortest path. We consider the simplest random graph model consisting entirely of closed triads and compare it to a network where the links are randomly rewired. We find that the giant component occurs at the same point—when the average node degree equals 1. However, past the phase transition, the giant component in the graph of closed triads grows more slowly than it does in the random network. We further examine the dependence of the average shortest path with the size of the network and find it to be almost identical for reachable pairs in both the triangle graph and the equivalent random network.

CHAPTER V

Information Diffusion in Citation Networks

5.1 Introduction

In the previous three chapters, some important structural features of the information sharing networks was studied. How these features impact the diffusion of information was analyzed. In this chapter, we present work about more direct and explicit relationships of information diffusion and network structures in paper citation networks and patent citation networks.

Information diffusion is the communication of knowledge over time among members of a social system. In order to analyze information diffusion, one needs to study the overall information flow and individual information cascades in the networks. Although much recent attention has been focused on new forms of collective content generation and filtering, such as blogs, wikis, and collaborative tagging systems, there is a well established social medium for aggregating and generating knowledge—published scholarly work. As researchers innovate, they not only publish new results, but also cite previous results and related work that their own innovations are based on. This creates a social ecology of knowledge—where information is shared and flows along co-authorship and citation ties.

Through their specialized organizations, activities, and publication venues, dis-

ciplines facilitate the frequent and timely dissemination of information. Within-discipline communication allows individuals to be exposed to research that is closest and most relevant to their own. Yet, there is a belief, reflected in many cross-disciplinary initiatives, both at the university and government levels, that knowledge flows between disciplines are not only beneficial, but are more likely to lead to innovative and groundbreaking research.

There is some evidence that interdisciplinary collaborations do lead to higher impact work. A study of scholarly articles in the UK found that papers whose coauthors are in different departments at the same university receive more citations than those authored in a single department, and those authored by individuals across different universities yield even more citations on average [51]. Multi-university collaborations that include a top tier-university were found to produce the highest impact papers [49]. Similarly, in the area of nanotechnology authors who have a diverse set of collaborators tend to write articles that have higher impact [92]. Interdisciplinarity aside, new collaborations between experienced authors are more likely to result in a publication in a high impact journal than new collaborations with an unseasoned author or repeat collaborations between the same two authors [46] . The argument is that merging ideas and expertise in a novel way will produce higher impact work. It has also been demonstrated that scholarly work in a range of fields and patents generated by larger teams of coauthors tends to have greater impact over time [124]. However, in the above studies examining author collaborations, there may be confounding factors. For example, successful authors may consequently have more opportunity to collaborate across departments and universities due to higher motivation or visibility.

In this chapter we aim to measure the impact of information flows from one field to

another more directly by tracing citations. Citations often, but not always, indicate that knowledge from one publication is being incorporated in another. Authors of the citing paper have found the other paper relevant, and more importantly, have usually, though not always [107], read it. Sometimes authors cite others where social norm or strategic positioning may encourage citation. Such behavior, if successful, would tend to reward citations within the same community or discipline, where one is targeting a publication. In the context of patents, inventors cite inventions that their own patent depends on or may be a substitute for.

We use as an indication of quality and impact of the work the number of citations a paper or patent receives normalized by the average number of citations received by all papers or patents in the same area and year [118]. This measure allows us to make a fair comparison between articles that may not have finished accumulating citations due to their recency, and to account for differences in size and publication cycle for different disciplines [114]. We take each individual citation as evidence of information flow, whether within a field or between fields.

The question we ask is simple: given the proximity in subject area between a citing publication (paper or patent) and cited publication, what is the impact of the citing publication? If cross-disciplinary information flows result in greater impact, one would see a negative correlation between proximity and impact. On the other hand, if it is within-discipline contributions that are most easily recognized and rewarded, one would observe a positive correlation.

5.2 Description of data sets

Our analysis uses two large data sets. The first, provided by JSTOR (Journal Store), has 1.98 million research articles in 1108 journals, classified into 47 disciplines,

roughly corresponding to 3 sets: arts and humanities, social sciences, and the natural sciences. Of those, there are 655,213 research articles citing 722,152 other articles within the dataset, for a total of 5,598,657 citations. These citations, limited to the cases where both the citing and cited articles are in the dataset, are a subset of the 23,451,235 citations made by the articles in total. Similarly, when measuring impact, we only count the number of citations from within the dataset. Although this could skew the observed raw citation counts toward disciplines that are better represented within the dataset, the normalization by discipline eliminates such biases. The patent data set contains all 5,529,055 patents filed between 1976 and 2006, and 2348 different categories with at least 1000 patents. There are 3,643,520 patents citing 2,382,334 others, for a total of 44,556,087 citations. We measure a patent's impact according to the number of other patents that cite it, normalized by the average number of citations for patents in the same year and class(es). The citation impact information is complete, since the dataset contains all subsequent patents.

5.3 Discipline proximity

Our analysis proceeds by examining each individual citation, the proximity of the disciplines of the citing and cited article for that citation, and the impact of the citing article. Intuitively, any individual citation will at most have a very weak impact on the success of a citing paper. It will only be one of possibly dozens of references made in an article or patent. Other factors, such as the publication venue and the reputation of the authors are more likely to contribute to the impact of the article than any individual citation the authors include. We nevertheless see a significant relationship between the interdisciplinarity of citations and the impact of the publication.

We assign disciplines to an article according to the JSTOR classification of the journal; approximately half of the journals are assigned to just one discipline, while the rest have multiple assigned disciplines. Each patent is assigned by a USPTO patent examiner to one or more categories according to the USPTO classification system. We quantify the proximity between disciplines by comparing the number of citations between any pair of disciplines relative to the rate of citation we would expect if the volume of inbound and outbound citations were the same, but the citations were allocated at random. If a citing or cited journal is classified into more than one discipline, a fractional citation is attributed to each discipline. We let n_{ij} be the actual number of citations from i to j, $n_{i\_}$ be the number of outbound citations from discipline i, $n_{\_j}$ be the number of inbound citations to discipline j, and n_T be the total number of citations. Then the expected number of citations, assuming indifference to one's own field and others, from field i to field j is $E[n_{ij}] = n_{i\_} \cdot n_{\_j}/n_T$. We define the directed proximity as a Z-score that tells us how many standard deviations above or below expected n_{ij} is:

$$Z_{ij} = \frac{n_{ij} - E[n_{ij}]}{\sqrt{E[n_{ij}]}}$$

Here we have used the observation that $n_T \gg n_{i\_}$ and $n_T \gg n_{\_j}$, and approximated the standard deviation by $\sqrt{E[n_{ij}]}$.

A high proximity between areas i and j indicates a strong tendency for papers or patents in area i to cite publications in area j. Figure 5.1 shows an information flow matrix of proximities by pairs of disciplines in JSTOR. Unsurprisingly, a discipline is most likely to cite itself. But one can also observe a tendency of the natural sciences to cite one another, while the natural and social sciences have fewer cross-citations. Furthermore, although the proximity from area i to area j is highly correlated with

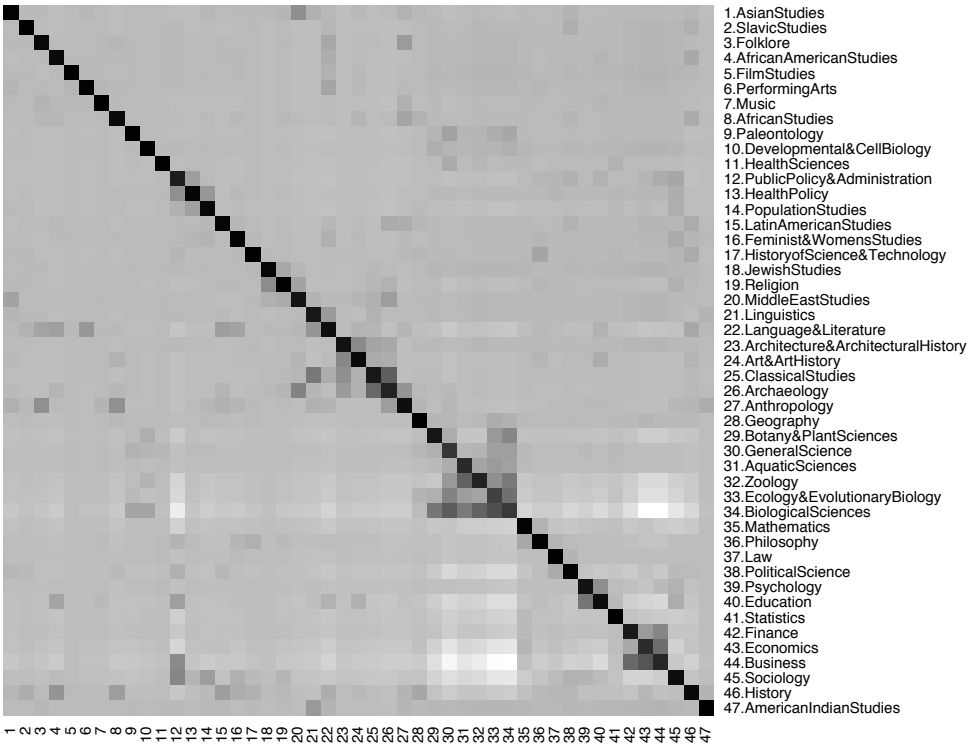


Figure 5.1: Information flow matrix for journals in the JSTOR database. The direction of information flow is from the column discipline to the row discipline, with Z_{ij}, the Z-score, corresponding to the i^{th} row and j^{th} column. Each entry is shaded according to a normalized Z-score representing whether the number of citations between disciplines is higher or lower than expected at random. Darker shading represents higher Z-scores. The diagonal represents citations within the same discipline.

proximity from j to i (with a Pearson correlation of 0.9675), the measure also captures any underlying asymmetry in citation patterns. Typically the more applied fields cite the more basic ones. Note that our measure is an aggregate over the entire lifetime of the journals included, and that previous time resolved measurements of information flow in chemisty-related fields have detected changes in flow as fields evolve [18].

In our aggregate sample, Finance cites Economics more often than Economics cites Finance. Statistics is more often cited by other fields than it cites them, with the exception of Mathematics. The areas of Zoology and Botany and Plant Sciences

cite the Biological Sciences more often than the Biological Sciences cite them. These asymmetries also reflect how unusual a citation is. A Biology paper citing a Statistics paper would be unusual, and might indicate the incorporation of a non-standard method. A Statistics paper citing a Biology paper would be even slightly more unusual, and might signal a motivation for the development of a novel method.

Figure 5.2 shows the information flow matrix for patents. For purposes of visualization, we have aggregated all citations according to 468 top level classifications (e.g., 029 corresponds to "metal working" while 901 corresponds to "robots"). We similarly observe a tendency of patents within the same subject classification to cite one another (patents are typically classified into several classes). Once more the proximity measure reveals asymmetries in information flow. For example, patents in category 623 "Prosthesis", which includes pacemakers for the heart, cite category 433 "Horology" more often than vice versa. Category 277, having to do with seals for a "joint or juncture" is more often cited by the categories corresponding to pumps and wells than it cites them. In general, those categories representing basic components and methods have a net surplus of citations, and include e.g., machine elements of mechanisms, gas separation, adhesives, stock material, and cryptography, among others. However, sometimes a category corresponding to a complex apparatus or process, such as 358 "Facsimile and static presentation processing" also has a net surplus of citations. This may occur when an invention matures and precedes other related inventions. The facsimile category is cited many times by other categories that developed later: television, computers, computer graphics, and interactive video.

In order to test the sensitivity of our results to our particular choice of proximity measure, in addition to the simple ratio of observed to expected citations, we also use the Jaccard coefficient for the sets of authors publishing in two areas. We select

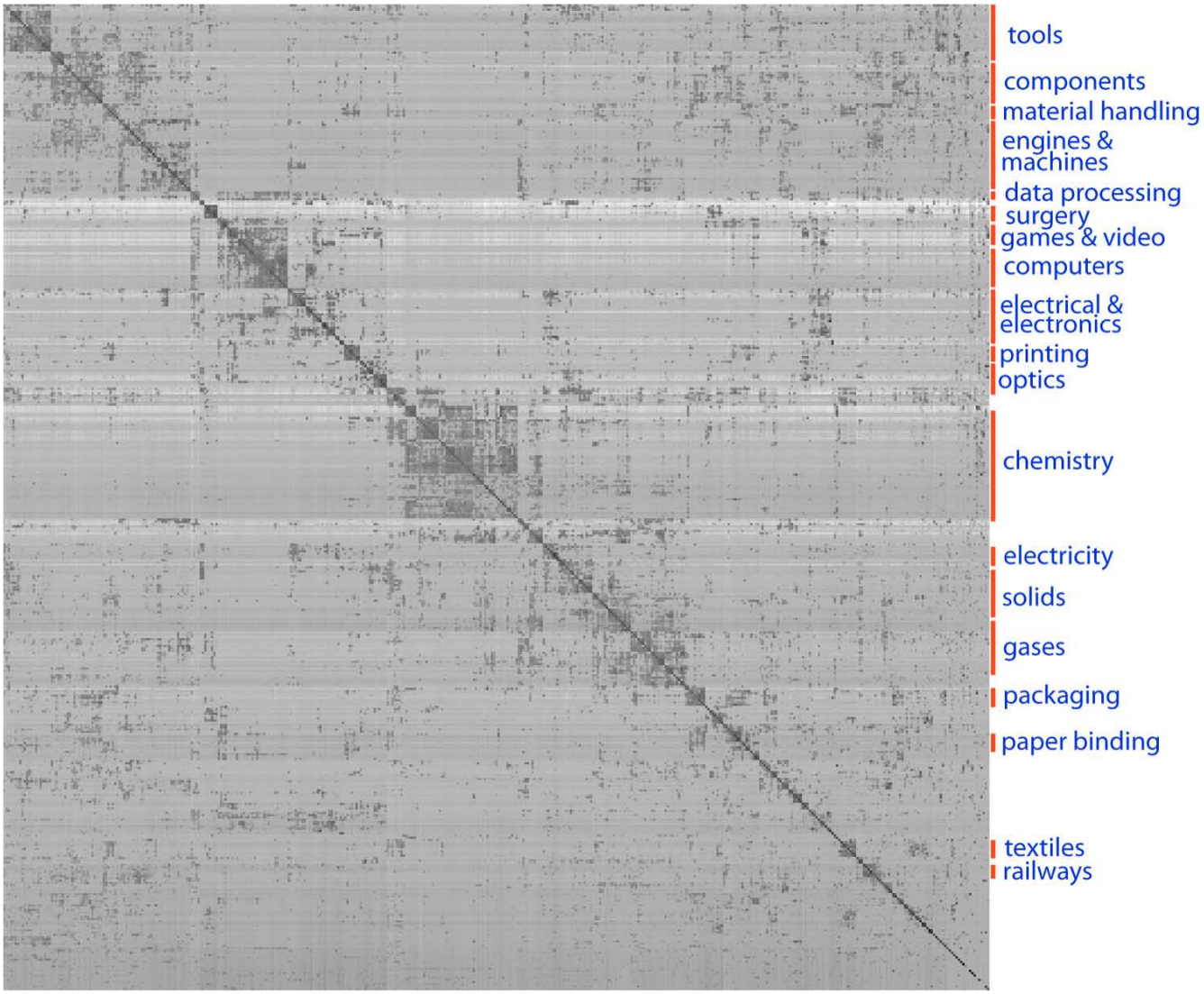


Figure 5.2: Information flow matrix for patents, with several related areas labeled.

the latter measure because it is very different from citation-based metrics, while still capturing proximity. An author could much more easily cite an unrelated area than they could directly contribute to it by publishing in that area's journals. In further contrast to the Z-score metric, the Jaccard coefficient is an undirected measure. Yet we still find our results, reported in Section 5.5, to be quantitatively and qualitatively consistent.

5.4 Impact of information flows

For every citing relationship, we measure the Spearman correlation between citation proximity and the impact of the citing publication. Citation proximity is simply Z_{ij}, where i is the area of the citing publication, and j is the area of the cited publication. If a paper or patent belongs to more than one area, the proximities are averaged. We sought to measure impact consistently across the diverse areas represented by our data sets. To that end, we measured impact (γ) as the the number of citations received by the citing publication, normalized by dividing by the average citation count of a publication in the same year and area(s).

We find that for the entire patent data set the correlation is positive ($\rho = 0.062^{***}$ [1]). The corresponding correlation for natural science papers in JSTOR is slightly negative with $\rho = -0.027^{***}$. However, one can also focus on publications with at least a given level of success. First, we omit the 40.03% of patents and 34.46% of natural science papers that were never cited within our datasets. After removing these zero-impact publications, the tendency of within-community citations to be rewarded is more significantly negative for both the natural science papers and patents: for patents, this correlation is -0.047^{***} and for natural science papers, the correlation is -0.072^{***}. This result suggests that a publication citing within its discipline is

[1] ***, **, and * denote significance at the < 0.001, < 0.01 and < 0.05 levels respectively.

more difficult to ignore altogether. However, given that a natural science publication or patent attracts at least some attention, there is a slight tendency for those that cite outside of their area to have higher impact.

To demonstrate that the result is not dependent simply on removing papers with no citations, we also slice the data according to percentile of impact, e.g., taking the bottom 30% and top 30%, and calculating correlations between citation proximity and impact separately for the top and bottom group. As Figure 5.3 shows, we consistently observe a negative correlation between citation proximity and impact for the higher impact group.

Figure 5.4 helps to explain why removing zero and low impact publications leaves a negative correlation between citation proximity and impact. By plotting mean proximity as a function of impact, we observe that both very low and very high impact papers tend on average to cite outside of their area more often. Since very low impact publications include many publications that cited outside of their discipline but failed to attract notice, we are left with the portion of cited publications where citing outside of ones discipline is positively correlated with impact. These results suggest that citing outside one's discipline is a gamble. While risking not being cited at all, publications that incorporate work from other disciplines tend to make more significant contributions.

Interestingly, the correlation between the interdisciplinarity of citations and the impact of a publication in the social sciences and humanities remains positive to neutral regardless of whether one includes or excludes zero citation publications. In the social sciences the correlation is 0.033^{***} when zero impact publications are included, and 0.040^{***} if they are excluded. The correlation for the entire set of humanity papers is 0.044^{***}, and -0.011 (not sig.) after removing papers with zero

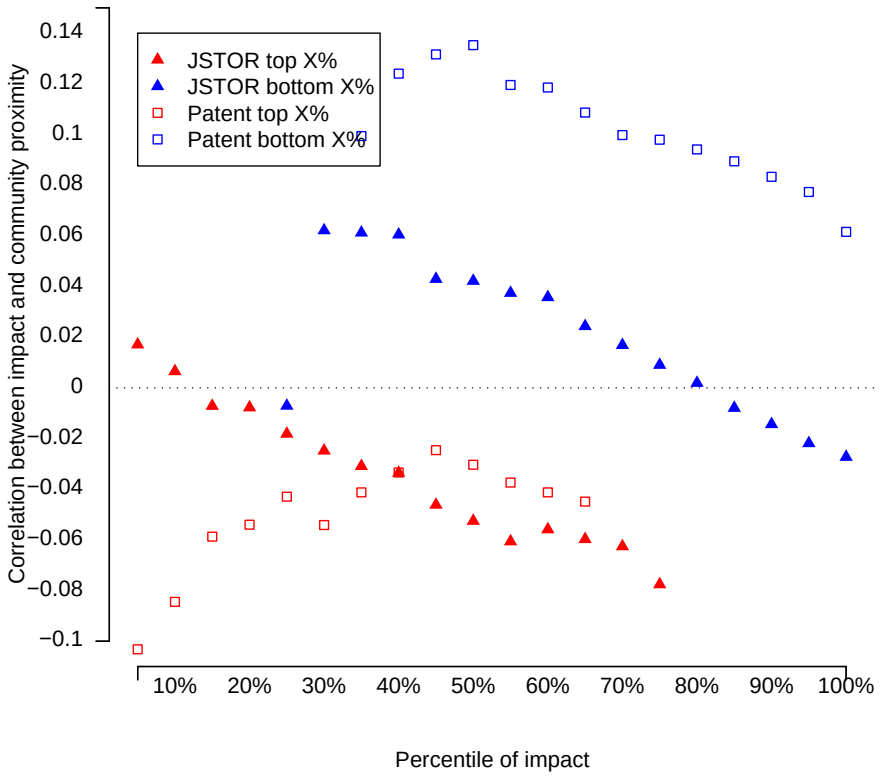


Figure 5.3: Correlations between proximity Z and impact γ, partitioned by percentile of impact. For example, at the 20% percentile, we show $\rho(Z, \gamma)$ for the bottom 20% of publications by their impact γ, and for the top 20% by γ. No correlations are shown for the bottom 10-20% of publications because they received no citations.

impact. That citing outside of one's discipline has different implications depending on whether one is a natural or social scientist is an interesting observation for further study.

In the above analysis, the correlation values are obtained individually by correlating the citation proximity and the impact of the citing publication for each citation pair. One can, however, also consider the average community proximity between a given publication and all of the publications it cites. Note that these averages are not always representative because many cited publications fall outside of our

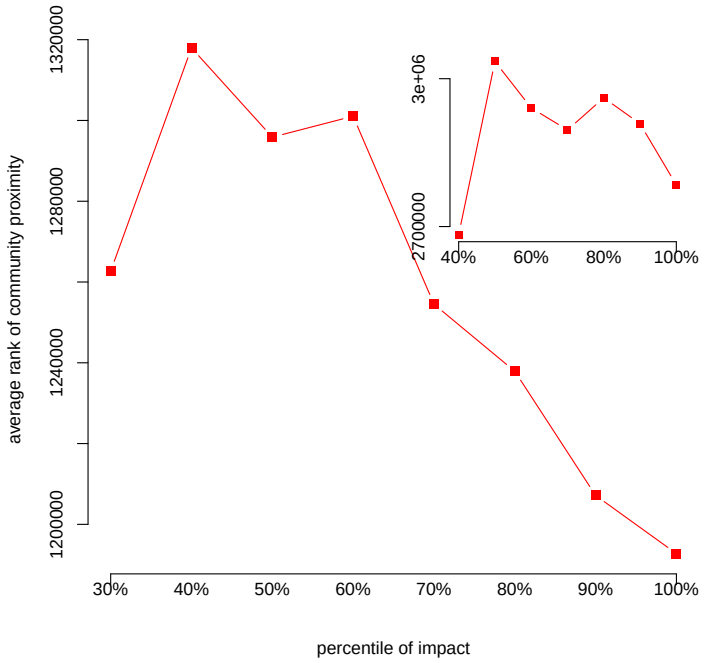


Figure 5.4: Average community proximity of citations by impact of citing article in JSTOR. The inset shows the average trend for patents .

datasets. Nevertheless, the correlation is 0.081^{***} for the entire set of patents, and -0.015^{***} for the set of patents having non-zero impact. For JSTOR, the correlations are -0.017^{***} and -0.028^{***} respectively for the set of natural science publications. These correlations are weaker, though consistent with the correlations obtained for individual citation pairs.

In order to interpret this result we should consider two scenarios for why an inter-community edge would appear. The first is that an author publishes in a venue outside their usual area, but cites work from their home area. It may be expected that their impact in the venue is diminished, possibly because the publication is of peripheral interest, or the Matthew effect [75] is absent, since the author has not already built up a reputation at that venue, and her work is less likely to be noticed. A second possibility is that an author who usually publishes in a given

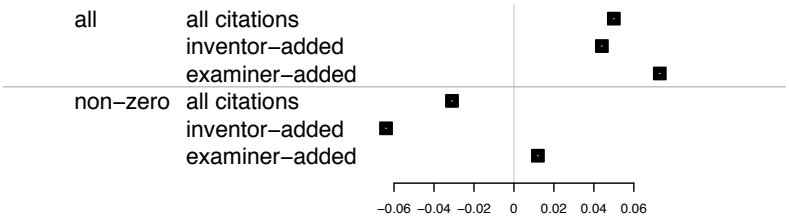


Figure 5.5: Correlations between citation proximity and impact, for patents published between 2000 and 2006, separated by whether the citation was added by an inventor or patent examiner.

venue draws upon another field in their work, sometimes by co-authoring directly with someone from another discipline [92]. One may expect such work to have potentially higher impact, since it is bringing in knowledge that could have greater novelty. Unlike journal publications where one may expect that impact will depend on both a suitably chosen venue and the innovativeness of the work, for patents there is only a single venue, the US patent office. Nevertheless, a patent's classification, determined by the patent office, affects its likelihood of being found by examiners and inventors searching the patent database.

Another way in which patents differ from journal articles is in the origin of the citations. As many as two thirds of all patent citations are added not by the inventors, but by the patent examiners, and it is therefore unlikely that such citations represent true knowledge flows [9]. Fortunately, since 2000, examiner-added citations are delineated from inventor-added ones. Already in the choice of patents to cite we find that examiners are more specialized in their citations than inventors; the average proximity for citations added by examiners is 213.471, compared to 155.572 for those added by inventors. Figure 5.5 shows that, unlike inventor added citations, examiner-added citations show a neutral to positive correlation for citing patents in proximate categories. This suggests that patent examiners may not only be biasing

citations to fall within categories, but when they do, the patent is more likely to receive citations.

Finally, we combine proximity with other variables which may influence the impact of the publication or patent. We include network properties of the citing and cited publications in the citation graph as well as the time of publication for both. We exclude variables such as publication venue and author since these themselves may be correlated with the likelihood of cross-disciplinary information flows. Table 5.1 gives the coefficients of the variables of the regression models. The dependent variables in these models are the impact of the citing paper of each citation pair after applying a Box-Cox transformation with an appropriate λ, i.e:

$$y' = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

Because of the extremely skewed distribution of the values of community proximity, we use their ranks instead of their normalized Z-score values. From Table 5.1, we see, consistent with results in Figure 5.3, that even controlling for other variables, cross-disciplinary citations correlate with higher impact for non-zero impact publications.

Furthermore, citing well-cited publications corresponds to receiving more citations, as does citing more recent publications. This is interesting in light of the recent finding that electronic access tends to make it easier to cite more recent and more influential papers [36]. Finally, citing many other publications positively correlates with receiving more citations. One might speculate that a publication that carefully acknowledges and builds upon a substantial body of previous work will itself be relevant to a wider range of future work.

Given the higher impact of information flows spanning disciplines, an important question one might ask is whether interdisciplinary citations have increased in recent

Table 5.1: Citing behavior and subsequent citations earned.

	US Patents		Natural science papers in JSTOR	
variable	all $_{(\lambda = 0.35)}$	> 0 cites $_{(\lambda = 0)}$	all $_{(\lambda = 0)}$	> 0 cites $_{(\lambda = -0.069)}$
$\log(\#\ \text{cited}_{\text{citing}} + 1)$	1.816e-01***	1.543e-01***	7.605e-01***	3.577e-01***
$\log(\#\ \text{citations}_{\text{cited}} + 1)$	1.470e-01***	1.047e-01***	2.635e-01***	9.971e-02***
citing year	-1.096e-02***	5.195e-05	-1.019e-02***	-7.828e-03***
year difference	-1.697e-02***	-1.092e-02***	-1.962e-02***	-7.209e-03***
proximity	**-5.873e-10*****	**-1.586e-08*****	**-1.743e-09*****	**-1.735e-08*****
R^2	0.0672	0.0534	0.1570	0.1018
citation pairs	2,841,279	2,683,726	2,110,965	1,729,298

$p < 0.05$(*), $p < 0.01$ (**), $p < 0.001$ (***)

years. Figure 5.6 shows the evolution of average community proximity over time for patents and for papers in JSTOR. We observe that the frequency of citations crossing communities among scholarly work has remained approximately constant over the past 100 years. For patents, we observe a mild increase in interdisciplinary citations from 1975 to 1990 and a sharper increase thereafter. This indicates that even though the amount of knowledge has been accumulating within each area, patent inventors and examiners are increasingly identifying and building upon relevant inventions in other areas. Note that our measures of proximity are based on the cumulative citation counts for the entire period of the datasets, which does not take into account variations in proximity between pairs of disciplines over time. Because of this, some pioneering papers that bring together disciplines before such cross-disciplinary research becomes common, may not be recognized in our analysis.

In summary, we quantified through a bibliometric analysis the effect of interdisciplinary information flows. We found that among patent inventions and natural science papers receiving one or more citations, those who cite across disciplines tend to garner more citations, indicating that cross-fertilization of ideas does often lead to significant impact.

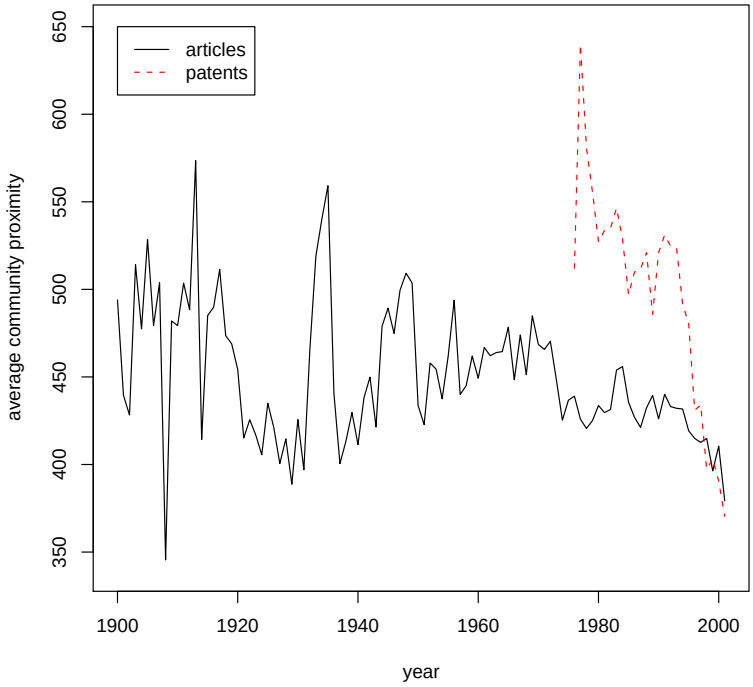


Figure 5.6: Average community proximity between communities over time.

5.5 Alternate definitions of proximity between communities

In addition to defining proximity in terms of citation frequency between areas/categories, one can also define it in terms of the author Jaccard coefficient p_{ij} that measures the ratio of the number of authors who publish in both areas to the number of authors who publish in either. Using the Jaccard coefficient has the feature of being 1 for all within community citations, and 0 for two areas that share no authors. In contrast, the community proximity measure has different weights for within-community citations because the Z-score measures how many more within-community citations than expected one observes, which varies by area. Therefore the Jaccard coefficient is able to treat all within-community citations equally.

We find generally good agreement between the two measures when correlated

against impact. For patents overall, the correlation drops to slightly negative using author overlap ($\rho = -0.008^{***}$), but is again significantly negative once the zero impact patents are removed ($\rho = -0.034^{***}$). Similarly for natural science articles in JSTOR, the non-zero impact articles have more significantly negative correlation ($\rho = -0.080^{***}$) compared with the overall correlation ($\rho = -0.037^{***}$). Once again, we have the result that inventions and natural science publications citing outside of their area tend to have slightly higher impact. For the humanities and social sciences, the correlations remain significantly positive both before and after excluding zero impact publications. Finally, the average p_{ij} for citations among patents and papers in JSTOR, shown in Figure 5.7 , is decreasing to constant, as was the case for the community proximity shown in Figure 5.6.

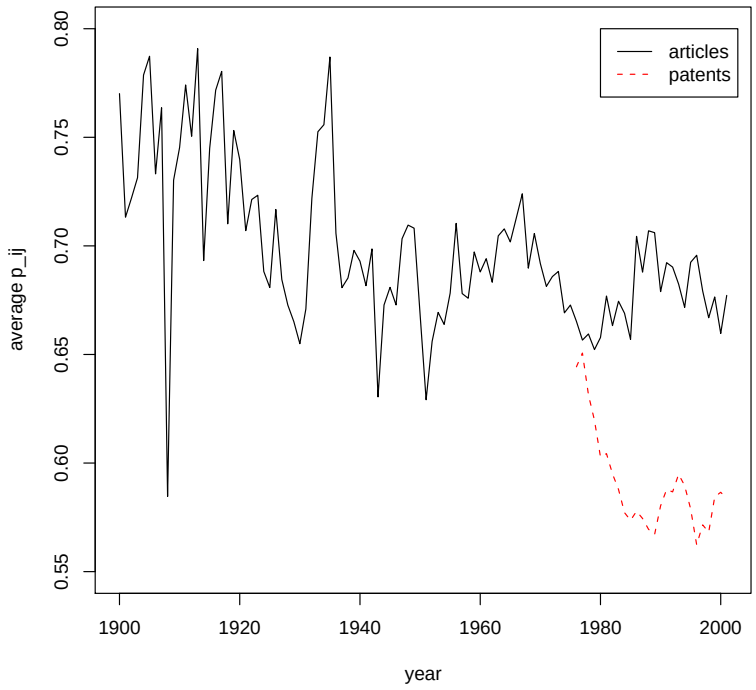


Figure 5.7: Average p_{ij} between communities over time.

5.6 Conclusions

In this chapter, we analyzed a very old, regimented, and established social medium for knowledge sharing in order to discover patterns of information flow with respect to community structure. There are interesting factors, relating to the citation graph, that correlate with the popularity a given publication will enjoy. Our particular interest is on the impact of a particular citation on the success of the citing work. Through intensive study of two large data sets, one spanning over a century of scholarly work in the natural sciences, social sciences and humanities, and one of a quarter century of United States patents, we find that for the most influential group of work, citations that occur crossing communities lead to a slightly higher number of direct citations. This is the evidence that the ideas across communities can lead to higher impact work.

We also examined citation patterns within a single discipline of scholarly publications, that of computer science publications. The datasets we study are two large digital libraries encompassing comprehensive scholarly articles primarily in computer science—the `ACM`[2] data set and the `CiteSeer` [3] data set [41]. The effects are less obvious, but still leave open the possibility that citation between disciplines could lead to high impact work.

[2] `http://portal.acm.org`
[3] `http://citeseer.ist.psu.edu`

CHAPTER VI

Information Diffusion in Online Forums

6.1 Introduction

In the previous chapter, we discussed the relationship between information flows and their subsequent impact in citation networks. In this chapter, we focus on the information diffusion patterns in online communities, i.e., the user grouping behavior in online forums. Online forums provide a unique type of social environment that enables people to share and access information freely. Users can either start new topics or leave comments in the threads of existing topics. Usually, an online forum has tens or hundreds of distinct boards or communities. These boards or communities group hundreds to thousands of threads of similar related topics together. Because of the huge numbers of users and the high dynamics of online forums, this type of environment has a rich complexity [43].

In this chapter, we focus mainly on three central questions:

1. What are the factors in online forums that potentially influence people's behavior in joining communities and how do they impact?

2. What are the relationships among these factors, i.e., which ones are more effective in predicting the user joining behavior, and which ones carry supplementary information?

3. What are the similarities and differences of user grouping behavior in forums of different types (such as news forums versus technology forums)?

By *a user joining a community* in a online forum, we mean the user posts at least once in the community. In this sense, "communities" are explicitly pre-defined, but the joining behavior is temporary and requires little effort. In the previous studies of information diffusion in other social environments, such as LiveJournal and DBLP [13], or a recommendation referral program run by a large retailer [67], the relationships between people are explicit and the actions taken require more commitment. However, the relationships or links in most forum networks are hidden and implicit—there are no well-defined links such as friendship or affiliations [43]. The most obvious relationship among users in online forums is the reply relationship between users. Instead of reflecting strong friendship, the reasons people are linked together by online replies may be because of common interests or different opinions [126, 42, 43].

In order to answer the first question, we analyze several features that can usually be obtained from a forum dataset. Our first discovery is that, despite the relative randomness and arbitrariness, the diffusion curve of influence from users of reply relationships has diffusion patterns similar to those in [13], although the reasons that people are linked together are very different. We also investigate the influence of the features associated with communities, which include the size of communities and the authority or the interestingness of the information in the communities. We find that their corresponding information diffusion curves show some strong regularities of user joining behavior as well, and these curves are very different from those of reply relationships. Furthermore, we analyze the effects of *similarity of users* on the communities they join, and find two users who communicate more frequently or have

more common friends are more likely to be in the same set of communities.

In order to answer the second question, we construct a bipartite graph, whose two sets of nodes are users and communities, to encompass all the features and their relationships in this problem. Based on the bipartite graph, we build a bipartite Markov Random Field (BiMRF) model to quantitatively evaluate how much each feature affects the grouping behavior in online forums, as well as their relationships with each other. BiMRF is a Markov random graph [37, 119] with edges and two-stars as its configuration, and incorporates the node-level features we have described as in a social selection model [94]. The most significant advantage of using the BiMRF model is that it can explicitly incorporate the dependency between different users' joining behavior, i.e., how a user's joining behavior is affected by her friends' joining behavior. In contrast, the decision trees as used in [13] cannot directly model such dependency. The results of this quantitative analysis shows that different features have different effectiveness in prediction in news forums versus technology forums. Together with results from the qualitative analysis, we are able to answer the third question. Our work also suggest that BiMRF models can be applied to analyze bipartite networks that are used to represent people and the common membership they belong to in general.

The findings discovered in this chapter are useful for improving and designing social network systems. Basically there are two important social functions in a social network system. One is how to recommend similar users, and the other one is how to recommend communities to users. The study of user grouping behavior reveals important features that have great impacts on how users join communities, and therefore provides valuable insights for social system owners to improve user experience. For example, a forum website can provide more social intelligence by

recommending top rated posts or large communities to users. It can also remind a user to pay more attention to other users who share similar interests with him. The findings related to the differences between news forums and technology forums also suggest that social systems should be designed with more considerations of diversified user intentions.

The rest of this chapter is structured as follows. Section 6.2 discusses some related work. Section 6.3 describes the datasets and network analysis results. Section 6.4 investigates dynamic features related to community-joining behavior. Section 6.5 presents the BiMRF model with quantitative analysis. Section 6.6 concludes this chapter.

6.2 Related work

The relationship between the user behavior and their social environment is the focus of a large body of work recently, such as [28, 43, 108]. The behavior of grouping is particularly interesting in social networks because it is closely related to the topic of *information diffusion* or *epidemics* [111].

The work in [13] also studies the human behavior of group formation. However, our work differs from it in the following aspects. First, the forum data, which have loose structures and hidden relationships, are different from the two social networks studied in [13]. The relationships between two users and between a user and a community in LiveJournal and DBLP require high commitment. For example, related neighbors have to be real friends in LiveJournal or co-authors in DBLP. In contrast, both user-user and user-community relationships in forums are much weaker because users do not have to exert much effort to have reply relationships with other users or participate in communities online. Second, in addition to the diffusion curves of

numbers of related users, our work also studies the diffusion curves of other forum features, the relationships between these features, and how the user behavior differs in news versus technology forums. Finally, instead of using decision trees [13], we use exponential random graph models, which can evaluate more complicated dependency features.

Another work studying user participation behavior is [64]. Instead of considering the relationship between users and communities, their target is to investigate the motivations of user participation on a social media site. The work [14] focuses on users who are heavily engaged in the group, and the behavioral differences between those users and ordinary users. They use a bipartite model to represent the user-group relationship; however, their model is to predict the "long-core" membership.

Users that do not participate publicly in online communities, i.e., people who lurk without posting, may be also interested in those communities. However, they are less positive in both activity and influence[83].

Exponential random graphs [95], which include the simplest Bernoulli random graph or Erdös-Renyi random graph model, Markov random graph [37, 119] and the recent developments [109, 96], have been extensively studied for social network analysis. Traditional use of random graph models is to discover structural statistics of networks, such as triangles and stars. Our work is an application of the homogeneous Markov random graph models [37, 119] with consideration of node-level attributes to give quantitative analysis of the forum data. BiMRF is a social selection model [94], in which individual users may change their joining behavior on the basis of the attributes of others.

6.3 Overview of the networks

In this section, we present an overview of the datasets and the bipartite networks of user-community relationships, as well as some structral features of these bipartite networks.

6.3.1 Datasets description

The datasets we study are from four online forums or online discussion platforms: Digg[1], Apple Discussions[2], Google Earth Community[3], and Honda-tech[4].

Digg is a news aggregator website, where users can submit news, videos, and pictures. In addition to that, users are able to lead discussions about the content that they are passionate about. All posted items, including news, images, videos and discussion comments can be rated by users by "digging" them. It is a platform on which people can provide content from anywhere on the web, and collectively determine the value of the information. The data we have crawled from Digg is from Oct., 2007 to Jul., 2008. It has 50 communities with topics of a great diversity. More than 200,000 users were active (i.e., posted at least once), and about 48,000 threads were built during that time period.

Unlike Digg, the other three forums focus on topics related to a specific product or technology. Apple Discussion is a platform mainly for Apple users seeking help, answering others' questions or exchanging opinions about Apple products. In our dataset, there are about 350,000 users and about the same number of threads in 331 different communities. The time window of this data ranges from 2001 to 2008. The forum of Google Earth holds discussions about the technology of Google Earth.

[1] http://digg.com
[2] http://discussions.apple.com
[3] http://bbs.keyhole.com/ubb/ubbthreads.php/Cat/0
[4] http://www.honda-tech.com

Our dataset has about 700,000 threads in 54 different communities, and 230,000 users were active from May, 2003 to June, 2008. A fraction of the posts in the Google Earth forum had ratings with them. Finally, Honda-tech is a forum for Honda customers to provide and exchange information and resources. It had 86,000 threads and about 45,000 users from 2001 to 2008. There were 63 communities in this forum. All of the four forums have explicit reply relationships in the datasets we have crawled.

6.3.2 User-community bipartite network

In social networks, bipartite networks or affiliation networks are bipartite graphs that are used to represent the people and the common memberships they belong to, such as the author-scientific article network, the actor-movie network [80]. In our problem, we define the user-community relationship as a bipartite graph: there is an edge between a user u and a community c, if and only if u has ever posted an article or a comment in c. Because little effort or commitment is required to post in online forums, the relationship between users and communities is not as strong as many other bipartite networks of user-membership. However, from the analysis of the bipartite networks, we are able to see some regularities of user joining patterns.

Table 6.1: Statistics about the bipartite networks.

Forum	Digg	Apple	Google Earth	Honda
User	212,635	349,066	231,976	45,718
Commu.	50	331	54	63
Edge	1,185,167	451,338	345,038	122,946
$\langle k_u \rangle$	5.57	1.29	1.49	2.69
$\langle k_c \rangle$	23703.34	1367.69	6389.59	1951.52
r	-0.2169	-0.0888	-0.2271	-0.0578

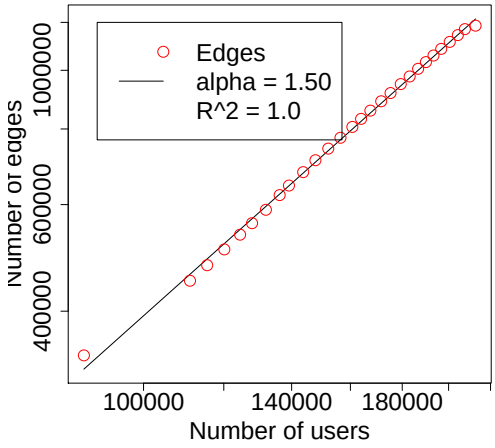
Table 6.1 gives a basic description of the user-community bipartite networks constructed from our forum datasets. $\langle k_u \rangle$ is the average number of communities a user joins, while $\langle k_c \rangle$ is the average number of users a community has. From the values of $\langle k_u \rangle$ and $\langle k_c \rangle$, we see the bipartite graph of Digg is much denser than the other

three. This shows that in news forums such as Digg, users are more likely to join multiple communities than in technology forums. We let r denote the value of assortativity, whose concept is defined as the preference of the nodes in a network to have edges with others that are similar under certain measurement [81]. Here we measure similarity with regard to degrees of nodes in the bipartite graph, and get the Pearson correlation coefficient between the degree of the users and the degree of the communities. We see that all four bipartite networks show negative values of r, which implies that in forums, less active users are more likely to join popular communities, while less popular communities are mostly occupied by active users.

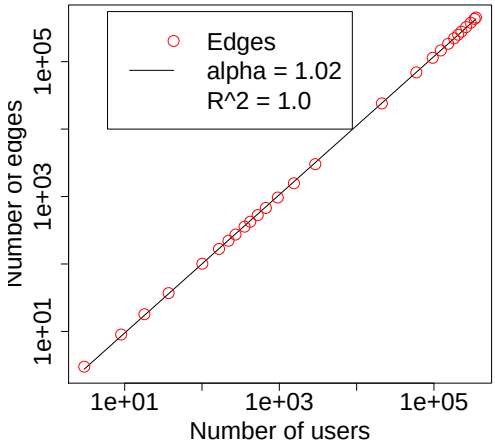
We then examine the growth of edges versus the growth of users in the bipartite networks of forums, by looking at whether their α of $e(t) \propto n(t)^{\alpha}$ follow the densification law [68]. In our bipartite networks, we assume that all communities existed since the beginning of our data availability, and that users start to join since the time they had their first post. From Figure 6.1, we see that the growth of edges is almost linear with respect to the numbers of nodes in the bipartite graphs of the four datasets. Being consistent with their low average degrees of users $\langle k_u \rangle$, this tells us that most users in the technology forums have much more focused interests and mostly stay in single communities. However, this is not the case for the forum of Digg, whose α is 1.5. In fact, we find that there are quite a few users who join almost all of the communities in this forum site.

6.4 Community membership

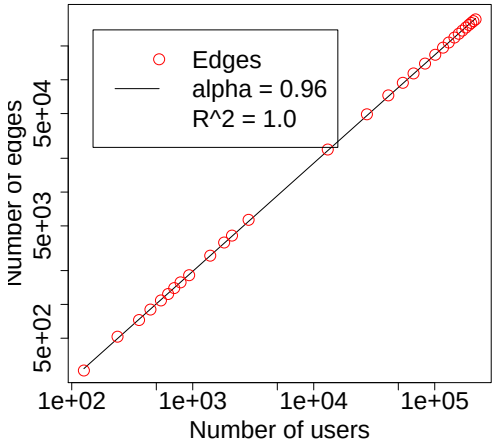
In the previous section, we analyzed some structural features of the online forum networks. In this section, we study the process of community joining behavior directly. In order to see the dynamics of user behavior, we divide the datasets into 30



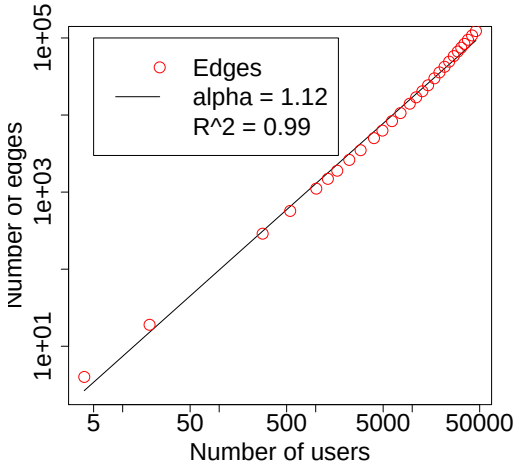
(a) Digg: $\alpha = 1.50$



(b) Apple: $\alpha = 1.02$



(c) Google Earth: $\alpha = 0.96$



(d) Honda: $\alpha = 1.12$

Figure 6.1: The growth of edges versus the growth of users in the bipartite networks.

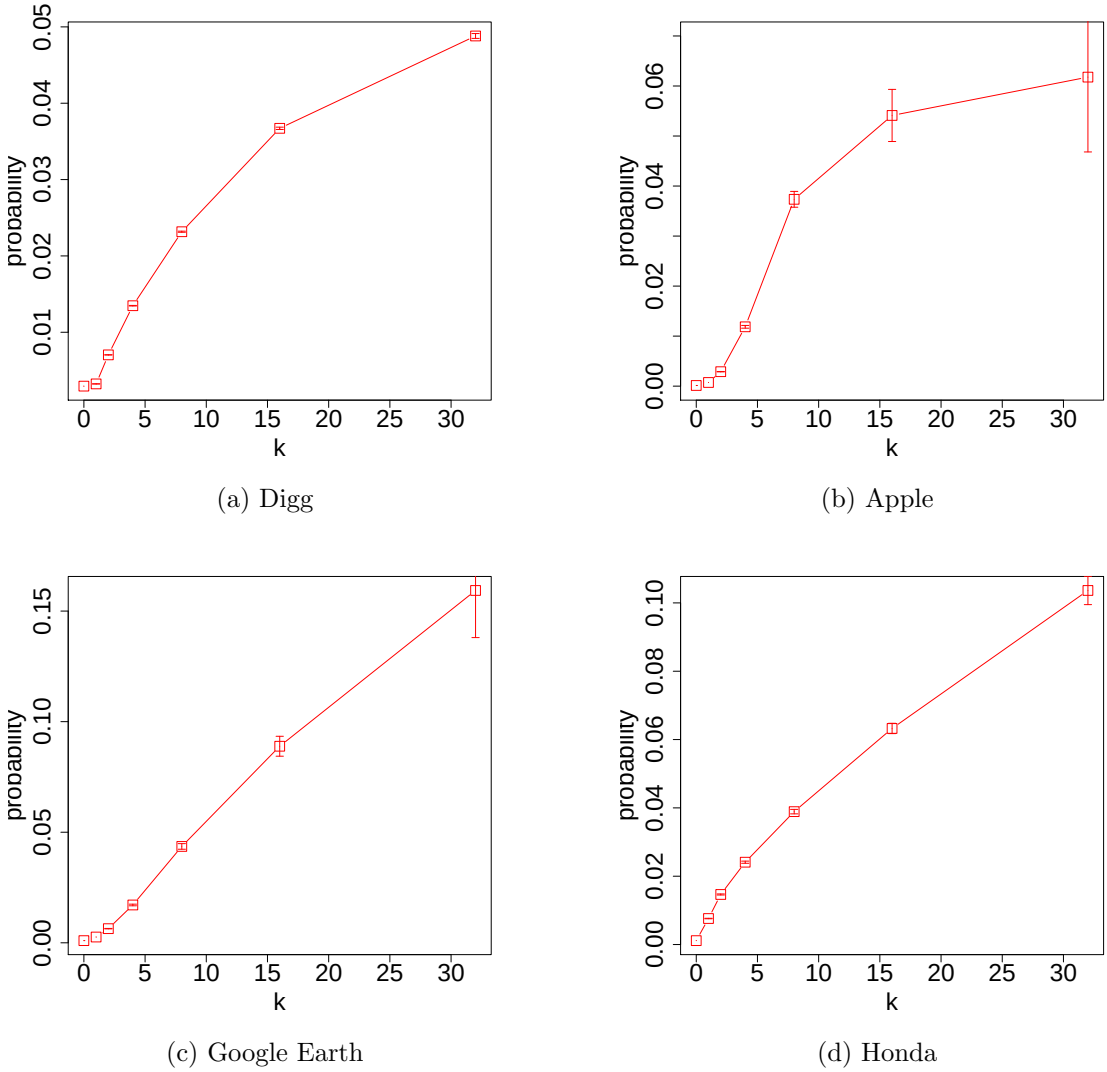
time snapshots. The diffusion curves we examine are the relationships between user joining behavior at time t and the related features at the previous time snapshot $t-1$. These curves show the change of joining probabilities as functions of different features associated with either users or communities. Moreover, we also study the correlations of user similarities and the communities they join.

6.4.1 Friends of reply relationship

We use this feature to describe how users are influenced by the numbers of neighbors with whom they have ever had any reply relationship. Although the reply relationship is not exactly the same as a real friendship, this is usually the most common and explicit user-user relationship that can be extracted from a forum dataset. In addition, as we will show, the reply relationship exhibits similar patterns in its diffusion curves as those of stronger relationships in other social networks, such as friendship or co-authorship in [13].

For every tuple (u, c, t) of user-community relationship at time t, we look at the reply friends of u who were active in c at the previous time snapshot $t-1$. We denote the number of such reply friends as k. By observing all the cases of whether u joins c with k reply friends at the previous time snapshot, we get the joining probability as a function of k. From Figure 6.2, we see that all four curves exhibit the *law of diminishing returns*. That is, the curves increase fast at the beginning, but more and more slowly towards the end. This is highly consistent with the observations of information diffusion in some other social networks [13, 67]. Moreover, the "S-shaped" behavior at $k = 0$, 1, 2 described in [13] is also observed in the three large datasets, Digg, Apple and Google Earth. The absence of this behavior in Honda may be because of the significantly smaller size of this dataset.

An alternative way to connect users is to let users in the same thread form a clique or a complete subgraph. However, this is a looser relationship than the reply relationship because users in the same thread may be interested in different aspects of the thread topic [126]. In fact, we also observe similar diffusion curves when considering the users in the same threads as 'friends', although the probability values are much lower. This is interesting since it suggests that, in many social networks,



(a) Digg

(b) Apple

(c) Google Earth

(d) Honda

Figure 6.2: The probability of a user joining a community in the forum as a function of the number of reply friend k who are active in that community at the previous time snapshot.

despite the diversity of 'friendships', their diffusion curves may have very similar patterns.

6.4.2 Community sizes

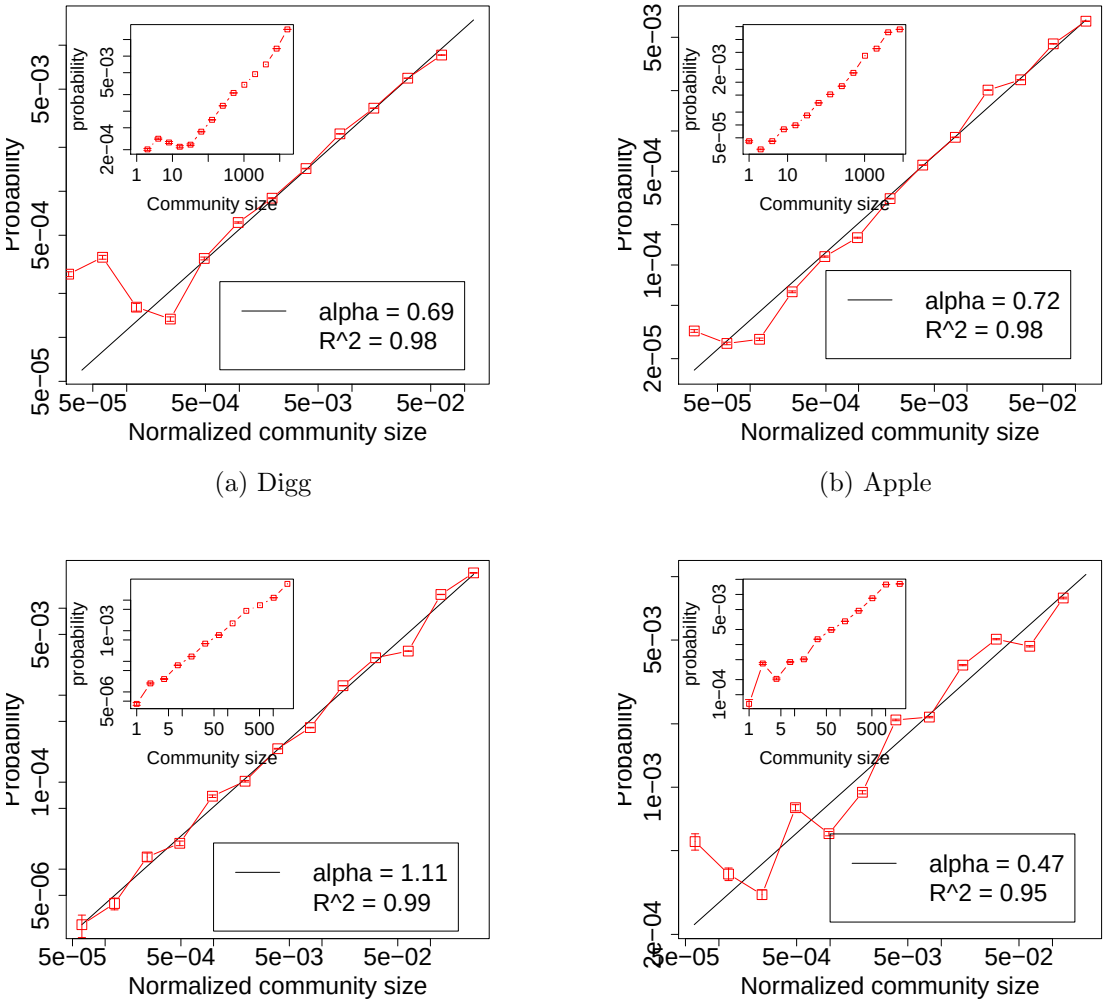
It is intuitive to expect that more popular information diffuses through the network at a faster pace. We examine this hypothesis in this part. We use *community size* as the measurement to quantify the 'popularity' of information.

By the *community size* at a time snapshot, we mean the number of users who have posted at least one article or comment in that community during that time snapshot. We call these users *active users*. The total sizes of all the communities at different time snapshots vary a lot, which may be because of both the limitation of the datasets and the effect of exponential growth of social communities. So we further normalize the community size over the sum of the sizes of all communities at that time snapshot.

Similar to the diffusion curves of the reply relationship, for every user-community tuple at each time snapshot t, we look at the normalized size of the community at time $t-1$, and get the user joining probability as a function of it. The curves are shown in Figure 6.3. The insets are the diffusion curves of absolute community size. From the figure, we can see that all the curves can be fitted by straight lines in the log-log scale. That is, if we use p and s to denote the probability of joining and normalized community size respectively, we have $p \propto s^{\alpha}$. We find that α is less than 1 in three of the figures, and larger but close to 1 in Google Earth. This tells us that the growth of the joining probability is sub-linear or linear with respect to the normalized community size.

6.4.3 Average ratings of top posts

Aside from the popularity of information, we are also interested in how the *authority* or *interestingness* of information impacts user behavior. Usually, in a social environment such as forums, the evaluation of the authority or interestingness of information is the result of the *wisdom of crowds*, since the ratings are the cumulative results of the users. In our datasets, Digg and Google Earth have rating systems, but their rating systems have some differences. First of all, the range of the ratings in Google Earth is from 0 to 5, while there is no upper bound of the ratings in Digg



(a) Digg

(b) Apple

(c) Google Earth

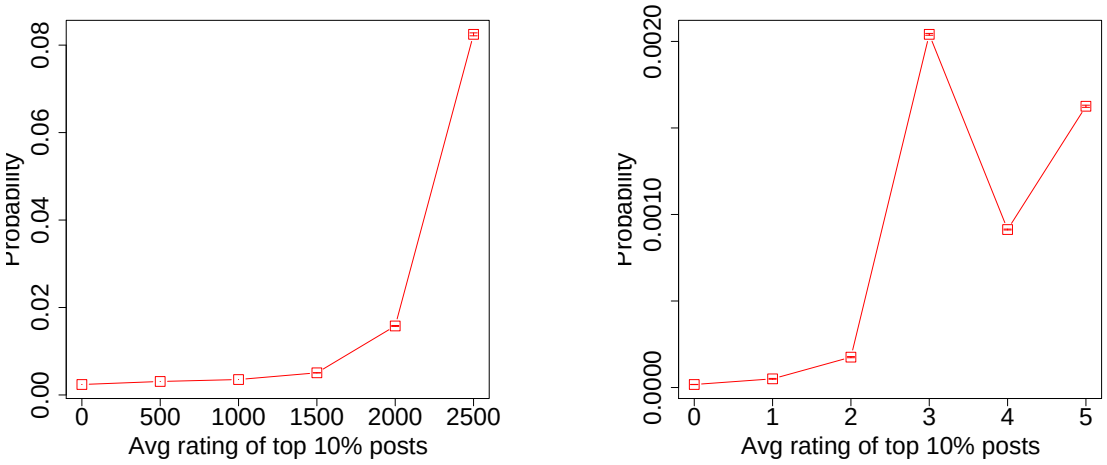
(d) Honda

Figure 6.3: The probability of a user joining a community in the forum as a function of the normalized community size at the previous time snapshot. The insets show the probability before normalization.

since those ratings are just the number of times a post has been "dug" by the users who like it. So the influence from the ratings in Digg may be confounded with the influence of community size, while Google Earth does not. Moreover, Digg allows ratings on starting posts as well as replies; while Google Earth only allows ratings on starting posts.

Users usually only see the ratings of starting posts before reading more of a thread.

So we only consider the ratings of starting posts as evidence for the authority of information. What is more, the data shows that in a community, the distribution of the ratings of starting posts is highly skewed, with most starting posts having very low scores and only a small fraction of them having high scores. Based on this fact, we choose the posts with top 10% ratings in each community at every time snapshot, and get the average of the ratings. Similar to the analysis of the previous two features, we plot the probability for users joining a community at a time snapshot as a function of the average rating of the top 10% posts in the community at the previous time snapshot. Figure 6.4 shows the resulting curves.



(a) Digg

(b) Google Earth

Figure 6.4: The probability of a user joining a community as a function of the average rating of the top 10% high rating posts in the community at the previous time snapshot.

It is interesting to see that there is a smoothly increasing curve for Digg, and the curve grows much faster after the average rating reaching a point around 2000. This curve shows a pattern that is called *critical mass*. On the other hand, the curve for Google Earth does not consistently increase as the one of Digg does. But still, the probability is much higher when the average rating is at 3, 4 or 5 than that of when

the average rating is at 0, 1, or 2. This difference between Digg and Google Earth might be due to people's different purposes in the two types of forums. In Google Earth, people are mainly seeking answers to their particular questions that may be only related to the topics in limited communities, so although the scores of the posts in the community matter, they do not have much difference after a threshold. However, the purposes people have for joining communities in Digg are more diverse. In addition, the front page of Digg enables users to read interesting topics without being aware of the communities they are in [64]. So increasing interestingness of the posts may be able to attract more users.

6.4.4 Similarities of users

In the previous part of this section, we have studied how certain features affect the probability of users joining communities. Those features are associated with either a single user or a single community. In this subsection, we analyze the features with dependency: if two users are 'similar' in a certain way, what is the correlation of the sets of communities they join?

To define the 'similarities' of users, two criteria are used. The first one is the number of times two users reply to each other's posts, normalized over the total number of articles or comments the two users have posted. The second one is the number of common friends that the two users have in the reply network, normalized over a half of the sum of the numbers of friends the two users have in total. For easy reference, we will name these two types of user similarity *frequency-user-similarity* and *triad-user-similarity* respectively. Each similarity measures takes values between 0 and 1. In order to get rid of noise introduced by trivial behavior, all users who only post once are ignored.

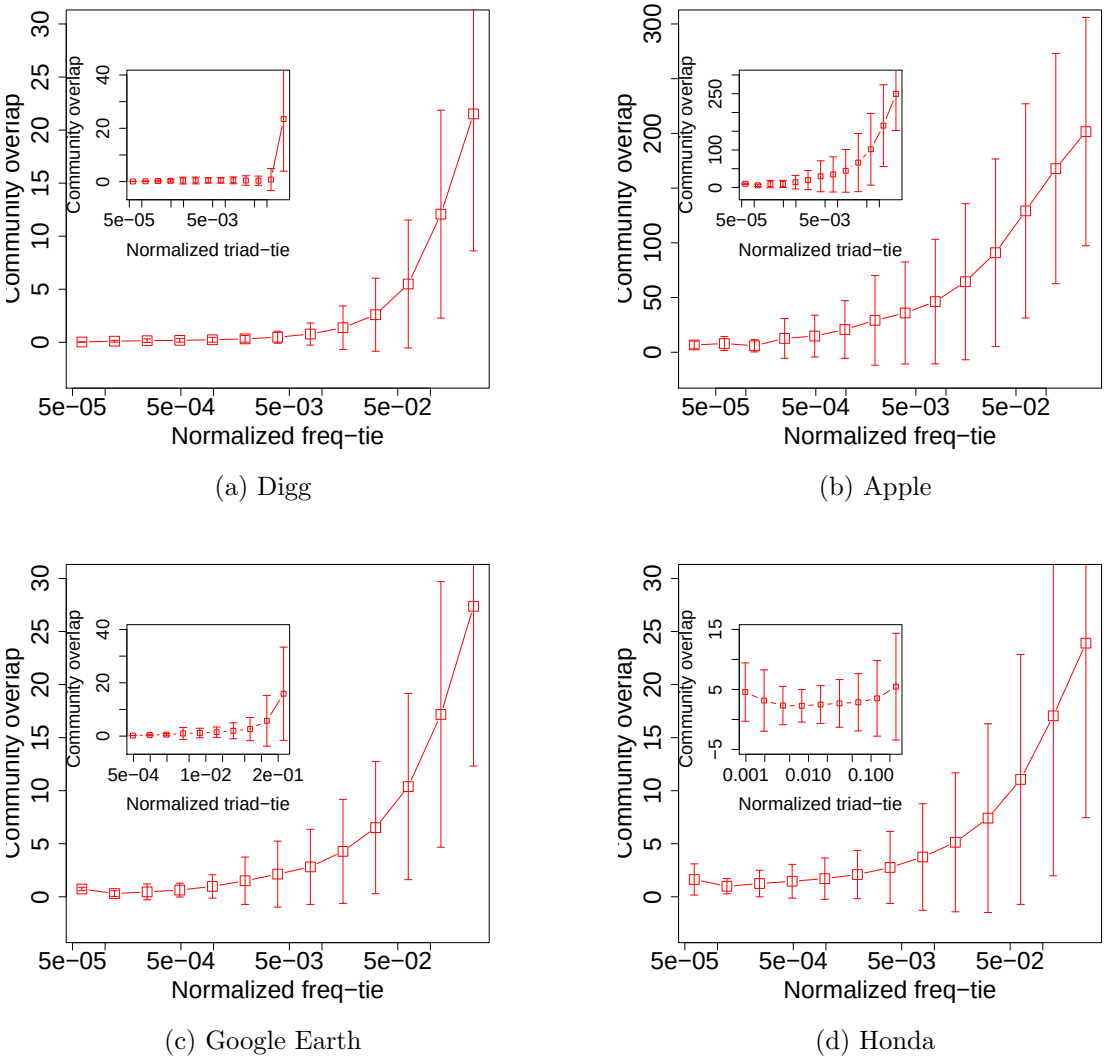
In order to know whether more similar users are more likely to join the same

communities, we compare their similarities versus the overlaps of communities they have joined. For two users u_1 and u_2, let the sets of the communities they have joined be S_1 and S_2, and the absolute overlap of their communities be $S_1 \cap S_2$. However, we need to account for the fact that some users may have little 'similarity' but large community overlap because they participate in almost all of the communities in a forum. So we normalize the absolute overlap by the expected overlap. The expected overlap can be obtained as $O_e = (|S_1| \cdot |S_2|)/(|S|)$, where S is the set of all communities in the forum. Then the normalized overlap O_n can be got by the equation: $O_n = \frac{(|S_1 \cap S_2| - O_e)}{O_e}$.

Figure 6.5 shows the relationships between user similarities and the normalized community overlaps. The correlation is positive for all forums and similarity measures, which means that more similar users are more likely to be in the same communities. We have to note that Figure 6.5 only shows the static correlations of user similarities and their community overlaps. This is different from the dynamic diffusion curves that we see in Figure 6.2 - 6.4. In fact, by computing the correlations between the user similarity at time $t-1$ and their community overlap at time t, we find they are neutrally correlated. This means two users either communicating more frequently or having more common reply friends at certain time are not more likely to join the same new communities in the following time snapshot. We will use a statistical model to further investigate this problem in Section 6.5.

6.4.5 Summary

In this section, we have shown how the community joining behavior is influenced by features associated with users and communities. The empirical diffusion curves show that these features are affecting human behavior in various ways. It is particularly interesting to see that the feature of reply friend has similar diffusion curves



(a) Digg

(b) Apple

(c) Google Earth

(d) Honda

Figure 6.5: The user similarities versus the community overlaps. The main plots use the communication frequency between users as the user similarity, and the insets use the number of common friends.

as those of real friend relationships in other types of social networks.

Moreover, we have analyzed the features of dependency. User-user similarities defined by their frequencies of communications and numbers of common friends are both positively correlated with the overlaps of the communities that the users have joined. However, there is no correlation between the user similarity and the sets of communities the users are going to join.

So far, we have examined the features separately. We will now consider them together to answer such questions as which feature best predicts user behavior and what correlations can be made with multiple features. We use a bipartite Markov random field model to study these problems.

6.5 Statistical user grouping model

In this section, we present a bipartite MRF (BiMRF) model, also known as a social selection random graph [94], to examine the quantitative effects of different features on the user grouping behavior in online forums. In addition to predicting user behavior based on the features observed, these models help reveal relationships between the features. Based on these relationships, we observe that the features have different effects in information diffusion in news and technology forums. As we shall see in Section 6.5.2, the advantage of using BiMRF models in our problem is that they can explicitly incorporate the dependency between related users' joining behavior, i.e., how a user's joining behavior affects her friends' joining behavior. The decision tree as used in [13] cannot explicitly model such dependency.

6.5.1 Bipartite markov random fields

In social network analysis, exponential random graph ($p^\star$) models have been extensively studied, including the simplest Bernolli random graph or the Erdös-Renyi model and the Markov random graph [37, 119] and its new specifications [109, 96]. In machine learning society, a Markov random graph is a Markov random field (MRF) with edges represented as random variables. In the sequel, we will obey this convention and point its connection to random graph models.

Based on the bipartite networks we have described in Section 6.3.2, we define the bipartite MRF (BiMRF) as follows. BiMRF is a bipartite graph and the vertices at

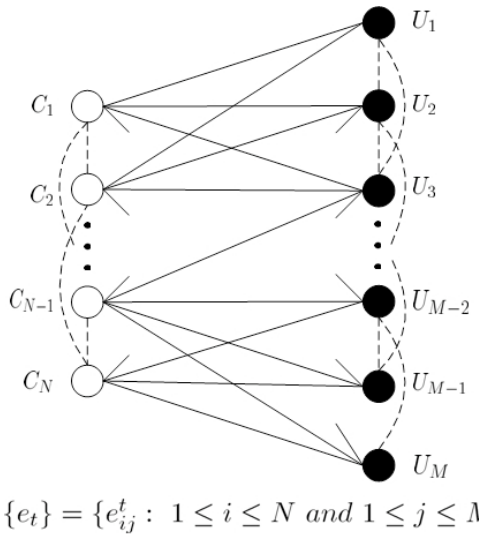


Figure 6.6: A bipartite MRF model with N communities and M users at time t. $\{e_t\}$ is an instance of the connections between users and communities at time t. The dashed edges are observed evidence.

one side are associated with the variables $U = \{U_i\}_{i=1}^{M}$ which represent users, and the vertices at the other side are associated with variables $C = \{C_j\}_{j=1}^{N}$ which represent communities. In the same spirit as the previous analysis, given the observed features, we treat the joining behavior at different time snapshots independently in the BiMRF model. Figure 6.6 shows the model's graph at time t. We will use the tuple (u, c, t) to denote the user-community relationship at time t. In our model, each user is a d-dimensional feature vector $u_i = [u_{i1}, \ldots, u_{id}]^\top$, of which the feature values can change over the time t as we have discussed in previous sections. Different users can be connected together, for example, if their similarity (by some measurement) is above some threshold. Each community c_j can have its features (e.g., community sizes) and can also connect to other communities if we have similarity defined between them, and their similarity is large enough. We will use O to denote all the observations, including users and their features, communities and their features, and connection structure of users and of communities. We introduce a set of random variables $E = \left\{E_{ij}^t : 1 \leq i \leq N \ , \ 1 \leq j \leq M \text{ and } 1 \leq t \leq T\right\}$, and each r.v. is an

Table 6.2: The two representative feature functions in BiMRF. *cs* denotes the features of *normalized community-size* and *us* denotes the two types of user similarity who are the same in defining feature functions.

Categories	Features	Feature functions
Singleton	*community-size*	$f_k^{cs}(e_{ij}^t = 1, c_i, u_j, t) = \begin{cases} 1 & \text{if } b^{-k} \leq \text{CommuSize}(c_i, t) \leq b^{-k+1} \\ 0 & \text{otherwise} \end{cases}$
Dependency	*user similarity (frequency or triad)*	$f_k^{us}(e_{ij}^t = 1, e_{il} = 1, c_i, u_j, u_l, t) = \begin{cases} 1 & \text{if } b^{-k} \leq \text{UserSim}(u_j, u_l, t) \leq b^{-k+1} \\ 0 & \text{otherwise} \end{cases}$

indicator: $e_{ij}^t = 1$ if the user u_i joins the community c_j at time t; otherwise, it is 0. Let $\{e\}$ denote an instance of the random variables E. By the basic theory of random fields [61], given the observations O, BiMRF defines a conditional distribution as follows:

$$p(\{e\}|O) = \frac{1}{Z(\mathbf{w})} \exp\Big(\sum_{k=1}^{K} w_k f_k(\{e\}, O)\Big)$$

where f_k are feature functions, which can be real or binary (here we assume they are binary, i.e., true or false), and w_k are their weights, which will be learned from a given training dataset. As we have mentioned, BiMRF treats the joining behavior at different time snapshots independently given the observed features. Thus, $p(\{e\}|O) = \prod_{t=1}^{T} p(\{e_t\}|O)$.

Since the dashed edges in Figure 6.6 are fixed and the probability $p(\{e\}|O)$ is defined on the connections between nodes on different sides, we call the model as a Bipartite MRF. A dashed edge is added if the similarity of the two users or the two communities at either side is above some threshold.

6.5.2 Feature function definition

We now define the feature functions for modeling user-community behavior. The features we use in BiMRF include three singleton features and two types of user similarity. The *singleton features* are those either associated with users or communities.

They are *reply-friend*, *normalized-community-size* and *top-post-rating* in our models, and will be denoted as rf, cs, tp respectively. The two *dependency features* are the two types of user similarity, i.e., *frequency-user-similarity* and *triad-user-similarity*, and we use us_f and us_t to denote them. We use the same bins as those used in the analysis of Section 6.4, i.e., we take linear-bin to define the feature function of *top-post-rating*, and the log-bin to define the other feature functions. Suppose the basis of the logarithm is b (e.g., 2 in our model). Two representative feature functions are defined in Table 6.2. Since f_k^{cs} are defined on an individual (u, c, t) tuple (i.e., a single joining event), we call these feature functions *singleton* feature functions; while f_k^{us} are defined on more than one joining events, thus we call these feature functions *dependency* feature functions. These dependency feature functions explicitly model the dependency between different users' joining behavior. In decision trees [13], however, such dependency cannot be directly modeled.

To avoid functions which appear sparsely in the datasets, we set an upper bound (e.g., 512) for the reply-friend feature and a lower bound (e.g., 2^{-19}) for the other four features. The features that are beyond this bound are defined in one feature function and will be treated the same in BiMRF.

These feature functions have a close connection to the configurations in Markov random graph [37, 119]. The singleton feature functions correspond to the dyad configuration and the dependency functions correspond to two-star configurations. In each type of configuration, we consider node-level attributes as in a social selection model [94].

6.5.3 Model fitting and testing

Model fitting is to learn the parameters from a given dataset. In this case, the data set is a pairing of observations and edges, i.e., $\mathcal{D} = \{\langle\{e\}, O\rangle\}$. The best model

to fit the data is the one with the maximum conditional likelihood: $\mathcal{L} = \log p(\{e\}|O)$. The optimization problem can be done with gradient ascent methods, such as the L-BFGS [72]. Since the probability is an exponential family distribution, the gradient is: $\frac{\partial \mathcal{L}}{\partial w_k} = E_{\hat{p}}[f_k] - E_p[f_k]$, where $E_p[.]$ is the expectation with respect to the model distribution $p(\{e\}|O)$, and $E_{\hat{p}}[.]$ is the expectation with respect to the empirical distribution on the given data corpus.

Without *dependency* feature functions, the Bipartite MRF models reduce to logistic regression models, also known as Bernoulli graphs in the social network literature. In this case, the model distribution, or the marginal probabilities as required in the objective function and its gradients, can be easily computed for each (u, c, t) independently.

With *dependency* feature functions, the BiMRF model is a homogeneous Markov random graph [37, 119] with two-star configurations. In each configuration, we consider node-level attributes as in a social selection model [94]. In a Markov random graph, the marginal probabilities on different edges, i.e., different (u, c, t) tuples, are coupled together. In other words, the event that a user joins a community at a particular time depends on the joining events of the related users or the communities at that time. Thus, we cannot compute the marginal probabilities of different edges independently.

For Markov random graphs, various estimation methods have been studied in social networks, such as the pseudo-likelihood method [113] and the Monte Carlo maximum likelihood estimation [120]. In this chapter, we use variational methods [50], which are among the most popular inference methods in the graphical model literature. The mean field approximation bears the form of pseudo-likelihood function [113]. But unlike the pseudo-likelihood method, mean field marginal probabilities

are computed iteratively using the coupled mean filed equations given initial values.

Mean field inference

To illustrate how mean field inference works in BiMRF, we use the user similarity feature as an example. The following derivations can be easily extended to other BiMRF models. The BiMRF model defines the following joint distribution:

$$p(\{e\}|O) \propto \exp\left\{ \sum_{k=1}^{K^{us}} w_k^{us} \sum_{ijlt} f_k^{us}(e_{ij}^t, e_{il}^t, c_i, u_j, u_l, t) \right\}$$

We define the factorized variational distribution $q(\{e\}|O) = \prod_{ijt} q(e_{ij}^t|O)$ as an approximation to the joint distribution. To find the best approximation $q^\star$, we minimize the KL-divergence: $KL(q(\{e\}|O)||p(\{e\}|O))$. The optimization problem can be solved by an alternating minimization method. Specifically, at each step we solve the problem with respect to only one marginal distribution $q(e_{ij}^t)$ and keep all others fixed. Then, we can get the following coupled mean field equations by using q_{ij}^t to denote $q(e_{ij}^t = 1|O)$:

$$q_{ij}^t \propto \exp\Big(\sum_{k=1}^{K^{us}} w_k^{us} \sum_l q_{il}^t f_k^{us}(e_{ij}^t = 1, e_{il}^t = 1, c_i, u_j, u_l, t) \Big)$$

These coupled mean field equations reflect our intuition that the event that user j joins community i at time t is dependent on whether other connected users l join the community at that particular time. We iteratively solve the coupled equations to get a fixed point solution, which gives the (approximate) marginal probabilities.

Table 6.3: Distributions of the number of related users on different datasets for frequency-user-similarity.

#Related Users	Digg	Google Earth	Apple	Honda
≤ 20	63.93	97.03	96.81	69.80
≤ 40	75.89	98.80	98.51	81.05
≤ 60	82.30	99.29	98.98	86.30
≤ 100	88.40	99.59	99.34	91.48

Table 6.4: Evaluation results of different BiMRF models on the four datasets.

BiMRF	**Digg**		**Google Earth**		**Apple**		**Honda**	
Models	ROCA	AP	ROCA	AP	ROCA	AP	ROCA	AP
$\{cs\}$	0.700	0.00536	0.860	0.00697	0.912	0.00296	0.833	0.00542
$\{rf\}$	0.718	0.00922	0.520	0.00128	0.522	0.00025	0.640	0.00743
$\{cs, rf\}$	0.800	0.01295	0.862	0.00738	0.913	0.00310	0.853	0.01257
$\{us_f\}$	0.442	0.00271	0.477	0.00188	0.473	0.00014	0.467	0.00147
$\{us_t\}$	0.474	0.00911	0.467	0.00235	0.467	0.00018	0.483	0.00179
$\{us_f, cs\}$	0.699	0.00540	0.861	0.00734	0.912	0.00296	0.831	0.00542
$\{us_t, cs\}$	0.705	0.00551	0.860	0.00698	0.912	0.00296	0.832	0.00536
$\{us_f, rf\}$	0.570	0.00362	0.545	0.00122	0.532	0.00015	0.561	0.00162
$\{us_t, rf\}$	0.703	0.00708	0.526	0.00117	0.531	0.00015	0.588	0.00179
$\{us_f, cs, rf\}$	0.796	0.01276	0.861	0.00744	0.899	0.00295	0.851	0.01248
$\{us_t, cs, rf\}$	0.800	0.01301	0.862	0.00724	0.906	0.00307	0.853	0.01177

Prediction

Given a learned model, we can do prediction on unseen (u_j, c_i, t) tuples and get the marginal probability that an edge exists $p(e_{ij}^t = 1|O)$. This is the probability that the user u_j joins the community c_i at time t. Since joining events are rare, the probabilities $p(e_{ij}^t = 1|O)$ are much smaller than 0.5. We cannot use a threshold (e.g., 0.5) to decide whether a user joins a community. Instead, we use the ordering metrics ROC Area (ROCA) and Average Precision (AP) to evaluate the goodness of the models. We evaluate the results of the features individually as well as with different combinations. In each experiment, we randomly sample 70 percent of the (u, c, t) tuples as training data and predict on the rest in each dataset.

An issue with regard to the models with user-similarity features is that we need to take care of the large number of related users as defined by user-similarity. For

example, for the frequency-user-similarity, the maximum number of related users on Digg is 43,269, 15,205 on Apple, 3740 on Google Earth, and 2236 on Honda. These large numbers will destabilize the computation when performing mean field inference. Fortunately, as shown in Table 6.3, for frequency-user-similarity, most of the users have small numbers of related users. The case of the triad-user-similarity is similar. Thus, we can use a pruning method to remove those rare users who have a large number of related users. In this experiment, we apply a simple strategy. We remove a user's user-similarity features if the number of her related users is larger than K, which is a pre-specified parameter. We set K at 20 in our experiments. We tried other parameters (e.g., 40 or 60), and the results do not change much.

6.5.4 Observations

Singleton features. In Section 6.4 we have seen the diffusion curves related to reply friends and community sizes, however, we cannot compare their effectiveness in predicting the user joining behavior from those curves. The BiMRF models help us do this. From the first two rows of Table 6.4, we see that for Google Earth, Apple, and Honda, the *community-size* feature predicts user joining behavior much better than *reply-friend* does. In particular, *reply-friend* has very little effect in Google Earth and Apple (their ROCA values are around 0.5). In contrast, *reply-friend* performs slightly better than *community-size* in Digg. Furthermore, by comparing the first three rows of Table 6.5, we see that although *top-post-rating* performs worse than *community-size* in Digg and Google Earth, it is better than *reply-friend* in Google Earth while worse than *reply-friend* in Digg.

These observations suggest that in the three technology forums, users' joining behavior correlates more closely with the features associated with communities, such as community sizes and average ratings of the top posts in the communities, rather

than the number of reply friends of users. On the other hand, in a news forum such as Digg, the user behavior has a stronger correlation with the number of their reply friends. The possible reasons for this difference are as follows. First, in both Google Earth and Apple, about 53% - 54% of users have only one post, while there are about 27% such users in Honda and 33% such users in Digg. This may explain the relatively poor performance of *reply-friend* in Google Earth and Apple, since there are large fractions of users who do not have any reply friends before joining any community, and they do not have any further activity after getting some reply friends. Second, from the low average degrees of users and the almost linear growth of edges versus the users in Section 6.3, we know that most users in the three technology forums like to stay in one community from the time they joined, i.e., they do not tend to switch their focus or expand their interests to different communities. In this way, the properties of the communities are more essential for users to decide which community to join at the very beginning, because no matter how many reply friends they gain, it is not likely for them to follow their reply friend to other communities. However, users do not have such focused interests in Digg, so their interests are more likely to change to other communities as their reply friends do.

Table 6.4 and 6.5 list the main results of the models using different features. Table 6.5 shows the results related to *top-post-rating*, which appears only in Digg and Google Earth. From the quantitative measures in these two tables, we make several observations regarding the features.

Dependency features. The results (the fourth and fifth rows of Table 6.4) of BiMRF models using two user similarities tell us that these dependency features perform poorly in predicting, e.g., their ROCA scores are all below 0.5. Note that although Figure 6.5 shows that there are positive correlations between the similarities

Table 6.5: Evaluation results of the top-post-rating, and user-similarity on Digg and Google Earth.

BiMRF Models	**Digg**		**Google Earth**	
	ROCA	AP	ROCA	AP
$\{tp\}$	0.639	0.00404	0.760	0.00229
$\{cs\}$	0.700	0.00536	0.860	0.00697
$\{rf\}$	0.718	0.00922	0.520	0.00128
$\{tp, cs\}$	0.708	0.00568	0.882	0.01040
$\{tp, rf\}$	0.774	0.01155	0.765	0.00250
$\{tp, cs, rf\}$	0.804	0.01371	0.884	0.01080
$\{tp, us_f\}$	0.642	0.00418	0.765	0.00236
$\{tp, us_t\}$	0.647	0.00454	0.761	0.00230
$\{tp, cs, rf, us_f\}$	0.802	0.01378	0.885	0.01044
$\{tp, cs, rf, us_t\}$	0.804	0.01375	0.883	0.01075

of users and the overlaps of the communities they belong to, those correlations are static and do not reflect the dynamic relationship of the users' similarities and their future joining behavior. And our analysis in Section 6.4 gets the neutral correlations between the user similarities in a time snapshot and the overlaps of communities they are going to join in the next time snapshot.

By a close examination of the joining probabilities predicted by the BiMRF models, we see that many (u, c, t) tuples have a probability larger than 0.1, which is much larger than the average joining probability in the datasets. This means that the BiMRF models with only the dependency features, which correspond to two-star configurations in Markov random graphs, are inadequate for the online forum data. But these models can be improved by incorporating node-level attributes, as shown by the results of BiMRF models with both dependency and singleton features in Table 6.4 and 6.5. This suggests that user-similarity has a weak effect on joining behavior in online forums, and thus adding the dependency features of user similarity does not help improve the performance. Finally, we must point out that the naïve mean field we are using in BiMRF makes a very strong independence assumption about the variational distribution q. This may give a poor approximation to the true

distribution. Extending to the generalized mean field [125], which incorporates more structural dependency in q, could be helpful to get a better approximation.

Feature combinations. By combining all the singleton features (as in the third row of Table 6.4 and the sixth row of Table 6.5), we see that the results are significantly better. This is especially true for Digg. Thus we can conclude that all these three singleton features carry supplementary information with each other, although in the three technology forums, *community-size* outperforms other two significantly.

6.6 Conclusions

In this chapter, we investigated the user participation behavior in diverse online forums. Our study of the structural features of their user-community bipartite networks suggest that, compared with news forums, users' interests in technology forums are more focused in single communities instead of crossing communities. Moreover, the diffusion curves show how the features of reply friends and some attributes associated with the communities have influence on community joining. Although a reply friendship is a much looser relationship [43], it has similar diffusion curves of *diminishing returns* as real friendship and co-authorship in [13]. Furthermore, the statistical BiMRF models present some interesting relationships among these features. In particular, reply friend and community attributes have about the same effectiveness in prediction in the news forum, while in the other three technology forums, features associated with communities are more effective in prediction. These features also provide supplementary information in our model. Finally, our analysis of two-star dependency social selection models suggests that the weak user-similarity features cannot fit the forum data well by themselves and adding node-level features can improve the fit.

As our analysis shows that user preference of information is tied with their related users in the past, and different types of information attract users in different ways, our work provides suggestions on building social systems, such as personalized recommendation systems [84]. Moreover, using the methodology presented in this chapter, more detailed studies can be conducted to evaluate other features that may affect users' social behaviors. For example, the user interactions in our study is based on their explicit reply information. In fact, similar analysis can be done based on more hidden behaivors such as browsing, which is known to website owners. They can then use the insights gained from such data to inform their recommendations to users who lurk without posting. Utilizing textual analysis in forum data, and investigating user behavior related to diffusion of discussion topics is also a future direction.

CHAPTER VII

Summary and Conclusions

7.1 Conclusions

This thesis shows that, in spite of the high complexity of information dynamics in various information sharing networks, the relationship of the structural features of the networks and the information diffusion among them have some strong regularities. Moreover, utilizing these regularities would help with further information search and management [86, 84].

We start with a very important and fundamental question—how much our observations are affected by the incompleteness or limited time windows of datasets we are studying. Usually information sharing networks are formed by a huge number of nodes, and these numbers are exponentially increasing. Thus, it is mostly impossible to get all the data to analyze the global properties of those large dynamic information sharing networks. Usually people use various sampling methods to collect a small fraction of the data to analyze the whole network. Then researchers face the question about the robustness of those analysis regarding the global features when incorporating different time durations and means in crawling the data. The work in Chapter II mainly focuses on answering this question in blogosphere. It shows that for the two different samples of blogosphere, BlogPulse and TREC, in spite of the

low overlap in their coverage and time durations of data collection, some topological features, both local and global, show great consistency. The chapter also shows that, as for the dynamic nature of the blogosphere, when the time duration of a crawl is extended, the features start to converge. This tells us that by having some fairly comprehensive samples of the blogosphere, one can start to obtain good estimates of the topological features of the whole space. We even consider the existence of noise in such networks, for example, the effects of the existence of splogs (i.e., spams of blogs) in the blogosphere. We found that splogs contribute a fair fraction of the total link volume in the blogosphere, and consequently affect the degree distributions greatly.

In addition to the analysis on the sets of vertices obtained from different time windows and different crawling methods, we study a special set of vertices in online information sharing networks in Chapter III—the important vertices, which are of the highest values under certain importance measures so long as they are far fewer in number than the vertices in the original networks. We find that the subgraphs induced by the important vertices are good approximations of the original networks in terms of the information transmission and communication among the important vertices. The empirical analysis of three real-world online networks shows that the important vertices are much more closely and densely connected to each other. They have significantly shorter pairwise paths, and their relative ranks are almost all highly correlated to their ranks in the original networks. This observation gives us the strong evidence that subgraphs induced by important vertices are effective for information transmission and communication in large networks, and are good representatives of the original networks in many aspects. The experimental results with different networks (either social or technological) and importance measures are consistent.

After exploring the regularities of the structures related to vertices of the information sharing networks in Chapters II and III, in Chapter IV, we study a special edge structure in online friendship networks—the strong ties, which are defined as edges belonging to closed triads. From the experimental study of two real-world social networks, we find that these online friendship networks are robust with respect to the removal of weak ties. There remains a giant component that is smaller but still occupies a majority of the graph, and the average shortest path changes modestly, which means strong ties are capable of transmitting or gathering information widely and effectively over the whole network. In addition, we consider a simple random graph model consisting entirely of closed triads and compare it to the corresponding Erdös-Renyi random graph. The theoretical random graph model also shows a low cost in terms of connectivity and diameters of the networks of strong ties.

In Chapter V and Chapter VI, we are interested in the relationship between information flows and the community structures of various networks as social media. We have analyzed two large data sets of citation networks in Chapter V—one is of research articles provided by JSTOR, and the other is of United States patents. We find that many publications went mostly unnoticed, while some garnered considerable attention. In the citation graph, there are interesting factors that are predictive of the popularity a given publication will enjoy. We find evidence that citations that occur across communities lead to slightly larger subsequent impact in citation networks of publications of natural sciences and patents.

In Chapter VI, we study the information diffusion patterns in communities of online forums. The diffusion patterns of user behavior in joining various communities and the feature factors associated with users or communities that influence such behavior are studied. Furthermore, we built Bipartite Markov Random Field

(BiMRF) models to help understand the relationships of these features, as well as the differences in their impact in different types of online forums.

7.2 Work in Perspective

This thesis is centered around the study of information diffusion in networks, and the associated network properties that can possibly affect it, such as network topology, community structure and temporal features. For the future, such study can be extended in the following two directions.

Online information search and related human behavior analysis. The famous set of experiments of the social psychologist Milgram revealed two important underlying mechanisms of human behavior in social networks of the 1960s: people were just a few steps apart in the global social network, and were able to propagate messages or do decentralized search efficiently by using local information available to them. It is interesting to investigate what mechanisms would be revealed if such a set of experiments were performed in this era, when various online social media such as Facebook, LinkedIn, Wikipedia are creating new social structures and ways in which people acquire and consume information. One would expect that the answer may be very different and much more complicated than 40 years ago, and reveal more interesting and surprising social phenomena as well. To explore the relationships between network structures and information search and propagation in these new forms of social media may be an expansion of the work in this thesis: whether these networks are evolving towards structural patterns that are effective in forwarding and processing information, and why the underlying regularities exist and how they operate.

Moreover, we have to be aware that nowadays, many kinds of collaborative online

media are rapidly emerging, and behaviors such as online search, posting questions online, etc. have been among the most popular and easiest ways for people to look for information. Thus, these information networks not only allow for information propagation, but can also serve as tools for people to find information and get questions answered more directly than ever before. Topics related to human behavior patterns with regard specific types of information provided by these social media, such as how people react to gossip information in online forums, how people seek for answers in expertise networks, the patterns of user interests in YouTube, etc. would be another interesting direction to extend the work in this thesis.

Combining research in other areas, such as databases, information retrieval, etc., with research in this direction would help build online systems that are more effective for people to retrieve information from and get answers that they desire. For example, understanding the user behavioral patterns related to information search in social content networks would help design systems that are able to return search results that present a good tradeoff between relevance and diversity.

Understanding the predictability of social systems and the individuals. Another interesting extension is to understand the predictability of social systems, as well as of the individuals and their immediate neighborhoods. Problems here include the evolution of the network structures, the formation of communities, the participation of people in activities, the prediction of new links, the popularity of new ideas or products over time, and so on. While many studies show that some social systems are evolving according to certain laws (such as the mechanisms of preferential attachment and shrinking diameters of networks), recent work by Salganik, Dodds and Watts raised the possibility that the outcomes of certain types of social systems may be inherently unpredictable [99]. Thus, it is intriguing to ask which social

processes are predictable based on the observations of their early stages, and which ones are inherently unpredictable.

In order to advance in this direction, first of all, we may investigate the effectiveness of various known techniques, such as those of random graph modeling, machine learning and statistical analysis, in predicting some specific problems in social systems. These specific problems can be the formation of friendship networks, the diversification of online communities, and the distributions of popularity of videos or music on YouTube. Second, it is also interesting to further investigate the predictability of local actions and interactions based on the global properties of the social systems that the individuals reside in. Finally, it is also interesting to investigate the factors, such as advertisements, peer influence, or infusion of information, that may increase or decrease the predictability of a system.

From the perspective of social science, this research direction would help people understand the social environments that they are in; from the perspective of systems and engineering, it would help construct more productive and successful online social environments. A deep understanding of social networks needs to draw on a number of different disciplines, including psychology, sociology, economics, information science, and areas such as data mining and machine learning in computer science. These areas offer complimentary approaches and techniques to study problems in social networks.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] L. A. Adamic, O. Buyukkokten, and E. Adar. A social network caught in the web. *First Monday*, 8(6), June 2003.

[2] Lada A. Adamic. The small world web. In S. Abiteboul and A.-M. Vercoustre, editors, *Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries, ECDL*, number 1696, pages 443–452. Springer-Verlag, 1999.

[3] Lada A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of LinkKDD-2005*, 2005.

[4] Eytan Adar, Li Zhang, Lada A. Adamic, and Rajan M. Lukose. Implicit structure and the dynamics of blogspace. In *WWE2006*. ACM Press, May 2004.

[5] Micah Adler and Michael Mitzenmacher. Towards compressing web graphs. In *DCC '01: Proceedings of the Data Compression Conference (DCC '01)*, page 203, Washington, DC, USA, 2001. IEEE Computer Society.

[6] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM Press.

[7] R. Albert, A. Barabsi, and H. Jeong. Scale-free characteristics of random networks: The topology of the world wide web. *Physica A*, 281:69–77, 2000.

[8] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. The diameter of the world wide web. *Nature*, 401:130, 1999.

[9] Juan Alcàcer and Michelle Gittelman. Patent citations as a measure of knowledge flows: The influence of examiner citations. *Review of Economics and Statistics*, 88(4):774–779, 2006.

[10] Chris Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, July 2006.

[11] M. Ángeles Serrano and Marián Boguñá. Tuning clustering in random networks with arbitrary degree distributions. *Physical Review E*, 72:036133, 2005.

[12] M. Ángeles Serrano and Marián Boguñá. Percolation and epidemic thresholds in clustered networks. `http://arxiv.org/abs/cond-mat/0603353`, 2006.

[13] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, New York, NY, USA, 2006. ACM Press.

[14] Lars Backstrom, Ravi Kumar, Cameron Marlow, Jasmine Novak, and Andrew Tomkins. Preferential behavior in online groups. In *WSDM*, 2008.

[15] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[16] Johan Bollen, Herbert Van de Sompel, Aric Hagberg, Luis Bettencourt, Ryan Chute, Marko A. Rodriguez, and Lyudmila Balakireva. Clickstream data yields high-resolution maps of science. *PLoS ONE*, 4(3):e4803, 03 2009.

[17] K. Borner. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, 101(suppl_ 1):5266–5273, 2004.

[18] Kevin W. Boyack, Katy Börner, and Richard Klavans. Mapping the structure and evolution of chemistry research. In *ISSI 2007*, pages 112–123, 2007.

[19] K.W. Boyack, R. Klavans, and K. Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.

[20] M. Brady. Blogging: personal participation in public knowledge-building on the web. *Participating in the knowledge society: Researchers beyond the university walls*, 2005.

[21] M Brady. Blogs: Motivations behind the phenomenon. *Chimera Working Paper*, (2006-17), 2006.

[22] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[23] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Comput. Networks*, 33(1-6):309–320, 2000.

[24] Z. Burda, J. Jurkiewicz, and A. Krzywicki. Network transitivity and matrix models. *Physical Review E*, 69:026106, 2004.

[25] D. Cartwright and F. Harrary. A generalizationgeneralization of heiders theory. *sychological Review*, 63:277–292, 1956.

[26] Damon Centola and Michael Macy. Complex contagion and the weakness of long ties. `ftp://hive.soc.cornell.edu/mwm14/webpage/WLT.pdf`.

[27] Vittoria Colizza, Alessandro Flammini, M. Angeles Serrano, and Alessandro Vespignani. Detecting rich-club ordering in complex networks. *NATURE PHYSICS*, 2:110, 2006.

[28] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *SIGKDD*, 2008.

[29] D.J. de Solla Price. Networks of Scientific Papers. *Science*, 149(3683):510–515, 1965.

[30] I. Derenyi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical Review Letters*, 94:160202, 2005.

[31] D. Dieks and H. Chang. Differences in Impact of Scientific Publications: Some Indices Derived from a Citation Analysis. *Social Studies of Science*, 6(2):247–267, 1976.

[32] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Large scale properties of the webgraph. *Eur. Phys. J. B*, 38(2):239–243, 2004.

[33] Debora Donato, Stefano Leonardi, Stefano Millozzi, and Panayiotis Tsaparas. Mining the inner structure of the web graph. In *WebDB*, pages 145–150, 2005.

[34] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. k-core organization of complex networks. *Physical Review Letters*, 96:040601, 2006.

[35] P. Erdos and A. Renyi. On random graphs I. *Publ. Math. Debrecen*, (6):290–297, 1959.

[36] James A. Evans. Electronic Publication and the Narrowing of Science and Scholarship. *Science*, 321(5887):395–399, 2008.

[37] Ove Frank and David Strauss. Markov graphs. *Journal of the American Statistical Association*, 81:832–842, 1986.

[38] M. Garey and D. Johnson. The rectilinear Steiner problem is NP-complete. *SIAM J. Appl. Math.*, 32(4):826–834, 1977.

[39] A. C. Gilbert and K. Levchenko. Compressing network graphs. In *LinkKDD*, 2004.

[40] E. N. Gilbert and H. O. Pollak. Steiner minimal trees. *SIAM J. Appl. Math.*, 16(1):1–29, 1968.

[41] C. Lee Giles. Citeseer: Past, present, and future. In *AWIC*, page 2, 2004.

[42] K.-I. Goh, Y.-H. Eom, H. Jeong, B. Kahng, and D. Kim. Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions. *Physical Review E*, 73(6):066123, 2006.

[43] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in slashdot. In *WWW*, 2008.

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

[45] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW '04*, pages 491–501, New York, NY, USA, 2004. ACM Press.

[46] R. Guimera, B. Uzzi, J. Spiro, and L.A.N. Amaral. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(5722):697–702, 2005.

[47] Tad Hogg and Lada A. Adamic. Enhancing reputation mechanisms via online social networks. In *Proceedings of the 5th ACM conference on Electronic Commerce*, pages 236–237, June 2004.

[48] J. Jonasson. The random cluster model on a general graph and a phase transition characterization of nonamenability, 1999.

[49] B.F. Jones, S. Wuchty, and B. Uzzi. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science*, 322(5905):1259, 2008.

[50] Michael I. Jordan, Zoubin Ghahramani, Tommis Jaakkola, and Lawrence K. Saul. *An introduction to variational methods for graphical models.* M. I. Jordan (Ed.), Learning in Graphical Models, Cambridge: MIT Press, Cambridge, MA, 1999.

[51] J.S. Katz and D. Hicks. How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, 40(3):541–554, 1997.

[52] J. M. Kleinberg. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS)*, page 14, 2001.

[53] Jon Kleinberg. The convergence of social and technological networks. *Communications of the ACM*, 51(11):66–72, 2008.

[54] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[55] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. The Web as a graph: Measurements, models and methods. *Lecture Notes in Computer Science*, 1627:1–17, 1999.

[56] Pranam Kolari, Akshay Java, and Tim Finin. Characterizing the Splogosphere. In *WWE '06*, May 2006.

[57] Gueorgi Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 435–443, New York, NY, USA, 2008. ACM.

[58] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *WWW '03*, pages 568–576, New York, NY, USA, 2003. ACM Press.

[59] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Structure and evolution of blogspace. *Commun. ACM*, 47(12):35–39, December 2004.

[60] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178, 2005.

[61] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[62] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical Review E*, 73:016102, 2006.

[63] Thomas Lento, Howard T. Welser, Lei Gu, and Marc Smith. The ties that blog: Examining the relationship between social ties and continued participation in the wallop weblogging system. In *WWE 2006*, 2006.

[64] K. Lerman. User participation in social media: Digg study. In *SMA07*, 2007.

[65] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429, 2007.

[66] Jure Leskovec, Susan Dumais, and Eric Horvitz. Web projections: learning from contextual subgraphs of the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 471–480, New York, NY, USA, 2007. ACM Press.

[67] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, New York, NY, USA, 2006. ACM Press.

[68] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: Densification laws, shrinking diamaters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, IL, 2005. ACM Press.

[69] Ping Li, Kenneth Church, and Trevor Hastie. A sketch-based sampling algorithm on sparse data. Technical report, Department of Statistics, Stanford University, 2006.

[70] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, August 2005.

[71] Yu-Ru Lin, Wen-Yen Chen, Xiaolin Shi, Richard Sia, Xiaodan Song, Yun Chi, Koji Hino, Sundaram Hari, Jun Tatemura, and Belle Tseng. The splog detection task and a solution based on temporal and link properties. *In TREC blog Track*, 2006.

[72] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, (45):503–528, 1989.

[73] C. Macdonald and I. Ounis. The TREC Blogs06 collection: Creating and analysing a blog test collection. Tech report (dcs), Dept of Computing Science, University of Glasgow, 2006.

[74] Christopher McCarty, Peter D. Killworth, H. Russell Bernard, Eugene Johnsen, and Gene A. Shelley. Comparing two methods for estimating network size. *Human Organization*, 60:28–39, 2000.

[75] R.K. Merton. The Matthew Effect in Science. *Science*, 159(3810):56–63, 1968.

[76] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.

[77] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, Oct 2002.

[78] M E J Newman. Properties of highly clustered networks. *Physical Review E*, 68:026121, 2003.

[79] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. In *Proceedings of the National Academy of Science*, volume 101, pages 5200–5205, 2004.

[80] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *the National Academy of Sciences*, 99:2566–2572, 2002.

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

[82] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, 2001.

[83] Blair Nonnecke, Dorine Andrews, and Jenny Preece. Non-public and public online community participation: Needs, attitudes and behavior. *Electronic Commerce Research*, 6(1):7–20, 2006.

[84] Jeffrey M. O'Brien. The race to create a 'smart' google. *FORTUNE Magazine*, 2006.

[85] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, May 2007.

[86] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[87] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 814-818:440–442, 2005.

[88] J. Park and MEJ Newman. Origin of degree correlations in the Internet and other networks. *Physical Review E*, 68(2):26112, 2003.

[89] Romualdo Pastor-Satorras, Alexei Vazquez, and Alessandro Vespignani. Dynamical and correlation properties of the internet. *Physical Review Letters*, 87:258701, 2001.

[90] David M. Pennock, Gary W. Flake, Steve Lawrence, Eric J. Glover, and C. Lee Giles. Winners don't take all: Characterizing the competition for links on the web. *PNAS*, 99(8):5207–5211, 2002.

[91] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *PNAS*, 101(9):2658–2663, 2004.

[92] Ismael Rafols and Martin Meyer. Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 2008.

[93] E.J. Rinia, T.N. Van Leeuwen, E.E.W. Bruins, H.G. Van Vuren, and A.F.J. Van Raan. Citation delay in interdisciplinary knowledge exchange. *Scientometrics*, 51(1):293–309, 2001.

[94] Garry Robins, Peter Elliott, and Philippa Pattison. Network models for social selection processes. *Social Networks*, 23:1–30, 2001.

[95] Garry Robins, Philippa Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph ($p^\star$) model for social networks. *Social Networks*, 29:173–191, 2007.

[96] Garry Robins, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. Recent development in exponential random graph ($p^\star$) model for social networks. *Social Networks*, 29:192–215, 2007.

[97] M. Rosvall and C.T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327, 2007.

[98] B. Ryan and N.C. Gross. The diffusion of hybrid corn in two iowa communities. *Rural Sociology*, 8:15–24, 1943.

[99] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

[100] X. Shi, B. Tseng, and L.A. Adamic. Information Diffusion in Computer Science Citation Networks. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009)*.

[101] X. Shi, B. Tseng, and L.A. Adamic. Looking at the Blogosphere Topology through Different Lenses. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, volume 1001, page 48109, 2007.

[102] Xiaolin Shi, Lada A. Adamic, and Martin J. Strauss. Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 378(1):33–47, May 2007.

[103] Xiaolin Shi, Lada A. Adamic, Belle Tseng, and Gavin S. Clarkson. The impact of boundary spanning scientific publications and patents.

[104] Xiaolin Shi, Matthew Bonner, Lada A. Adamic, and Anna C. Gilbert. The very small world of the well-connected. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 61–70, New York, NY, USA, 2008. ACM.

[105] Xiaolin Shi, Jun Zhu, Rui Cai, and Lei Zhang. User grouping behavior in online forums. In *Proceedings of the fifteenth ACM SIGKDD conference on Knowledge Discovery and Data Mining*, 2009.

[106] C. Shirky. Power laws, weblogs, and inequality. In *"Networks, Economics, and Culture"*. Aula, Helsinki, Finland, 2003.

[107] MV Simkin and VP Roychowdhury. Read Before You Cite! *Complex Systems*, 14:269–274, 2003.

[108] Parag Singla and Matthew Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW*, 2008.

[109] Tom A. B. Snijders, Garry L. Robins, and Mark S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, pages 99–153, 2006.

[110] Xiaodan Song, Belle L. Tseng, Ching-Yung Lin, and Ming-Ting Sun. Personalized recommendation driven by information flow. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 509–516, New York, NY, USA, 2006. ACM Press.

[111] David Strang and Sarah A. Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 24:265–290, 1998.

[112] David Strauss. On a general class of models for interaction. *SIAM Rev.*, 28(4):513–527, 1986.

[113] David Strauss and Michael Ikeda. Pseudo-likelihood estimation for social networks. *Journal of the American Statistical Association*, 85:204–212, 1990.

[114] M.J. Stringer, M. Sales-Pardo, and L.A.N. Amaral. Effectiveness of Journal Ranking Schemes as a Tool for Locating Information. *PLoS ONE*, 3(2), 2008.

[115] M. P. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *PNAS*, 102(12):4221–4224, March 2005.

[116] Torsten Suel and Jun Yuan. Compressing the graph structure of the web. In *Data Compression Conference*, pages 213–222, 2001.

[117] Belle L. Tseng, Junichi Tatemura, and Yi Wu. Tomographic clustering to visualize blog communities as mountain views. In *WWE 2006*, May 2006.

[118] Jose M. Valderas;, R. Alexander Bentley;, Ralf Buckley;, K. Brad; Wray;, Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. Why Do Team-Authored Papers Get Cited More? *Science*, 317(5844):1496b–1498, 2007.

[119] S. Wasserman and Philippa Pattison. Logit models and logistic regressions for social networks, I. an introduction to exponential random graphs and ($p^\star$). *Psychometrika*, 62:401–425, 1996.

[120] S. Wasserman and Garry Robins. An introduction to random graphs, dependence graphs, and $p^\star$. *Models and Methods in Social Network Analysis*, pages 148–161, 2005.

[121] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 2002.

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

[123] Andrew Y. Wu, Michael Garland, and Jiawei Han. Mining scale-free networks using geodesic clustering. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 719–724, New York, NY, USA, 2004. ACM Press.

[124] S. Wuchty, B.F. Jones, and B. Uzzi. The Increasing Dominance of Teams in Production of Knowledge. *Science*, 316(5827):1036, 2007.

[125] Eric P. Xing, Michael I. Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. In *UAI*, 2003.

[126] Kou Zhongbao and Zhang Changshui. Reply networks on a bulletin board system. *Phys. Rev. E*, 67(3):036117, 2003.

[127] S. Zhou and R. J Mondragon. The Rich-Club Phenomenon In The Internet Topology. *IEEE Commun. Lett.*, 8:180–182, 2004.