UNIVERSITY OF SOUTHAMPTON

Supporting Meaningful Social Networks

by

Yongjian Huang

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

15$^{\text{th}}$ June, 2009

Recent years have seen exponential growth of social network sites (SNSs) such as Friend-ster, MySpace and Facebook. SNSs flatten the real-world social network by making personal information and social structure visible to users outside the ego-centric networks. They provide a new basis of trust and credibility upon the Internet and Web infrastructure for users to communicate and share information. For the vast majority of social networks, it takes only a few clicks to befriend other members. People's dynamic ever-changing real-world connections are translated to *static links* which, once formed, are permanent – thus entailing zero maintenance. The existence of static links as public exhibition of private connections causes the problem of *friendship inflation*, which refers to the online practice that users will usually acquire much more "friends" on SNSs than they can actually maintain in the real world. There is mounting evidence both in social science and statistical analysis to support the idea that there have been an inflated number of digital friendship connections on most SNSs. The theory of friendship inflation is also evidenced by our nearly 3-year observation on Facebook users in the University of Southampton.

Friendship inflation can devalue the social graph and eventually lead to the decline of a social network site. From *Sixdegrees.com* to *Facebook.com*, there have been rise and fall of many social networks. We argue that friendship inflation is one of the main forces driving this move. Despite the gravity of the issue, there is surprisingly little academic research carried out to address the problems. The thesis proposes a novel algorithm, called *ActiveLink*, to identify meaningful online social connections. The innovation of the algorithm lies in the combination of *preferential attachment* and *assortativity*. The algorithm can identify long-range connections which may not be captured by simple reciprocity algorithms. We have tested the key ideas of the algorithms on the data set of 22,553 Facebook users in the network of University of Southampton. To better support the development of SNSs, we discuss an SNS model called RealSpace, a social network architecture based on active links. The system introduces three other algorithms: *social connectivity*, *proximity index* and *community structure detection*. Finally, we look at the problems relating to improving the network model and social network systems.

This thesis is dedicated to my father, a healthy non-smoker, who had been battling lung cancer for nearly two years before he passed away in January, 2009.

Contents

List of Figures

List of Tables

Acknowledgements

I would firstly like to thank my supervisor Nigel Shadbolt for his wise advice, careful guidance and useful connections. In particular, Nigel's advice on the theory of complex network is the turning point of my research on social network. Learning how to research and produce publications is much easier with a supervisor who has as much experience and enthusiasm as Nigel.

I would also like to thank my second supervisor Nicholas Gibbins for his advice and comments, particularly in the writing up of my final thesis.

Wider thanks are also due to people who are willing to give their University email addresses to me, for making it possible to experiment with Facebook.

Chapter 1

Introduction

Recent years have seen exponential growth of social network sites (SNSs) such as Facebook, MySpace and Friendster, which have attracted hundreds of millions of Internet users over last few years. Figure 1.1 shows how the traffics of visiting social networks grow from 2002 to 2007. In particular, Friendster was launched in 2002 and gained huge popularity in 2004. MySpace was founded in 2003 and is the third most popular site in the US only behind Yahoo and Google. Facebook was founded in 2004 and has the largest number of registered users in the college. As of this writing, some statistics suggest that there are about 114.6 million users on MySpace, 123.9 million on Facebook[1] and 65 million on Friendster[2]. It is estimated that there are hundreds of SNSs, with various technological capabilities, supporting a wide range of interests and practices[19].

1.1 The Scope of Research

We define social network sites as websites that allow users to traverse others' social network of more or less institutionalised relationships of mutual recognition. The concept of relationship is an elastic term without clear demarcation. In this thesis, the institutionalised relationships mainly refer to relatives, classmates, colleagues and other friendships that are mutually acquired and recognised. In practice, most social networks provide profile services for users to present themselves and offer many other services such as friend feeds, messaging, blogging, photo and video sharing, groups and forums. The capability of navigating through the social network is unique to SNSs, in contrast to traditional computer-mediated communication (CMC) tools such as emails and instant messengers that mainly facilitate private one-to-one communications. Note that some social media such as blogging software, Twitter, del.icio.us, Youtube and Internet forums share some features with SNSs, yet while these Web applications offer a lot of

---

[1]http://www.comscore.com
[2]http://www.facebook.com/press/info.php?statistics
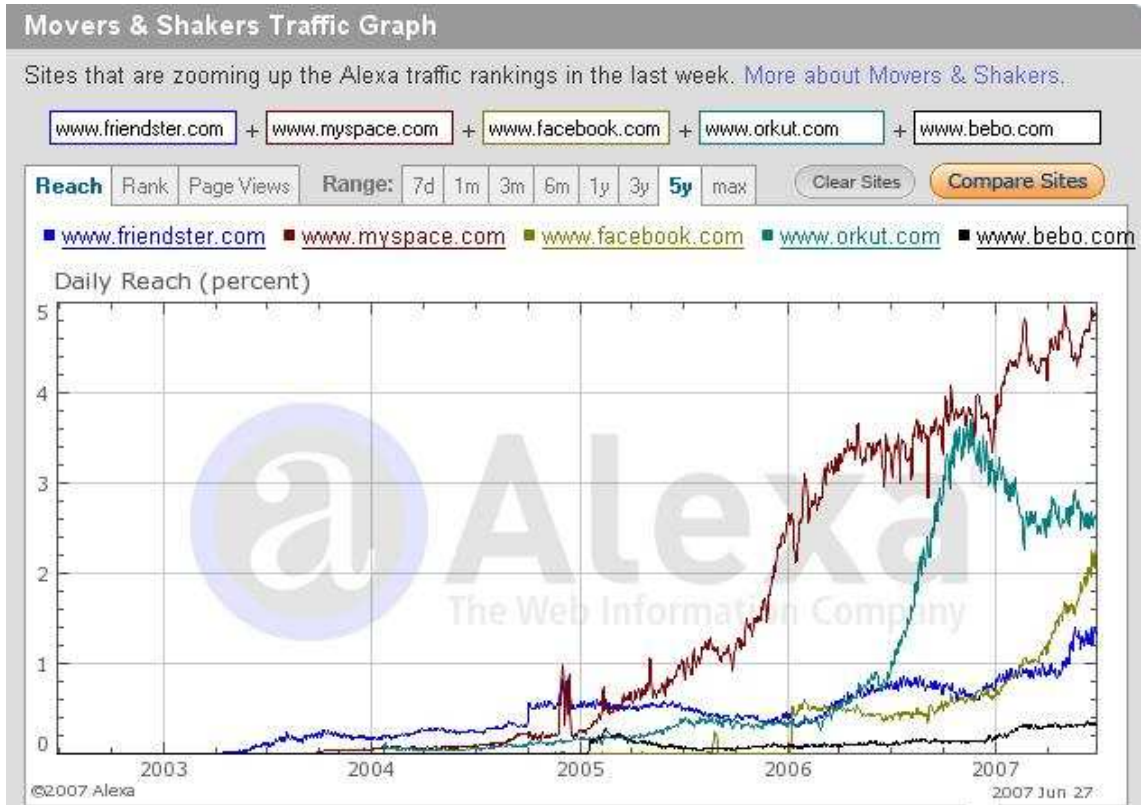
FIGURE 1.1: The Traffic of Friendster(2002), MySpace(2003), Facebook(2004), Orkut(2004) and Bebo(2005) from 2002 to 2007

information from user generated content (UGC), they do not focus on social relationships between the users. They generally lack of friend feeds as presented on most social network sites. Therefore, they are beyond the scope of this thesis. It should be noted that some scholars may well regard this type of websites as social network services.

The research of SNSs is still in its infancy, which can be seen by various interpretations of the acronyms. For instance, it is not uncommon for SNS to be interpreted as "social networking site" or "social networking service". The word "networking" means to initiate new relationships with strangers, which is one of the important motivations for using SNS. However, as stated earlier, we primarily focus on the "network" aspect of SNS rather than "networking". The word "service" implies a broader category of applications, with website service being only one of them. We prefer the term "sites" to reflect the fact that most SNSs are based on the Web platform. These terms are usually used exchangeably both in academia and industry.

One of our main objectives of this research is to study the social graph that lays the foundation of SNS[43]. The topology and structure of the network will affect people's behaviours and activities on the micro level, which in return affect the further development of the network on the macro level[61]. Given the large scale and evolving characteristics of social network, we are particularly interested in the problem of how to represent the

real-world social network correctly. The accurate representation of the real-world social network is essential, not only because users can navigate the network via meaningful connections but also because it can facilitate network analysis and any algorithm based on the network structure. The thesis introduces some network algorithms to enhance reputation and trust of individuals by leveraging the social graph. How to explore and navigate the network more efficiently will also be discussed.

1.2 Motivation of the Research

There are increasing academic interests in the research of social networks from a wide range of disciplines. Our research is mainly motivated by four aspects: (1) SNS is a new communication tool; (2) SNS is the online "bank" of social capital. It represents a good old-fashioned social networking paradigm based on existing real-world social relationships; (3) It advances the establishment of social Web where information and knowledge can be distributed at the right time to the right people with an enhanced layer of trust and security; (6) SNS can benefit the development of large-scale multi-agent systems by leveraging social intelligence.

1.2.1 Communication Media

First and foremost, SNS is a new communication tool. Social networks are capable to provide asynchronous communications through onsite messaging, public wall posting, etc. Many SNSs also offer synchronous communication by introducing instant messaging. The format of communication can come in many ways, from plain text to rich multimedia. Unlike one-to-one communication tools such as emails and instant messengers, users only need to publish their information once and their friends will be notified instantly. For example, this can be done on Facebook by using NewsFeed. Technologies of this type, according to Facebook, allow people to consume content from their friends and stay in touch with the content that is being shared. Users can subscribe to people and events they are interested in so as to receive the latest updates. Technologies such as these, according to Facebook, Some social networks provide support for mobile SMS so that users can access the sites by texting. The vast majority of social network sites are Web-based and therefore easier to reach more users than traditional communication software. This has been demonstrated by how SNS members use the social networks. One survey on Facebook, for example, indicates that the most use of SNSs was to keep in touch with friends from high school and find out more about a person they had met with offline[79]. Given its rich functionalities and user-friendliness, SNS is a powerful tool for massive grooming.

1.2.2   Social Capital

Social capital broadly refers to the resources accumulated through the relationships among people[27]. SNS holds the information about people's social relationships. SNSs provide Profiles for users to present themselves to other members. These Profiles contain information about personal identity. SNSs allow people to connect to their friends and make these connections visible to other users. These social connections are stored in the database of the social network sites. Users can carry out many social activities on the SNSs such as messaging, blogging, uploading photos and videos, commenting and public wall posting. SNSs can provide both asynchronous communication through instant messaging and synchronous communication through private messaging and public wall posts. Various communication channels, which are based on the ubiquitous Web platform, make it much convenient to socialise with other members. Some SNSs such as Facebook provide services like news feed which can aggregate most of users' activities and report to their friends. These online activities are also "deposited" on the social network sites. Users can manage their friends by using tools provided by the sites. Many SNSs impose permission control of this information according to users' preferences and privacy settings. Social network sites are online "bank" of social capital in the real world. Users can traverse their chains of friends and make new contacts through mutual connections. The method greatly reduces the time and effort for establishing trustful relationships. They help users to build up and maintain their social capital, which has a strong influence on business, economics, organisational behaviour, political science, public health and sociology. SNSs' contribution to increase of social capital has been confirmed by a fair amount of recent research[41][42][119][82][127].

1.2.3   Social Web

Social network services mirror people's real-world relationships on the cyberspace. They provide a new platform for people to share information and communicate. In May 2007, having already opened some APIs to third party developers, Facebook revealed its programming infrastructure and declared that the site was going to serve as a platform for programmers to develop applications, in the same sense that programmers can develop applications and software on the computer platform and the Web platform[3]. In November 2007, Google released a set of APIs for Web-based social network applications, which are dubbed "OpenSocial" and have been supported by a number of social network sites including Bebo, Friendster, LinkedIn, Mixi, MySpace and Orkut[4]. In May 2008, Google launched Friends Connect project, which aims to deliver social features to every website[5]. As a response, Facebook launched its similar service, Facebook Connect, in

---

[3]Facebook Platform Launches: http://developers.facebook.com/news.php?blog=1&story=21
[4]http://www.google.com/intl/en/press/pressrel/opensocial.html
[5]http://www.google.com/friendconnect/

July 2008. The development of standards and technologies make SNSs more ubiquitous and accessible on the Web. More importantly, it incubates a new platform for software development.
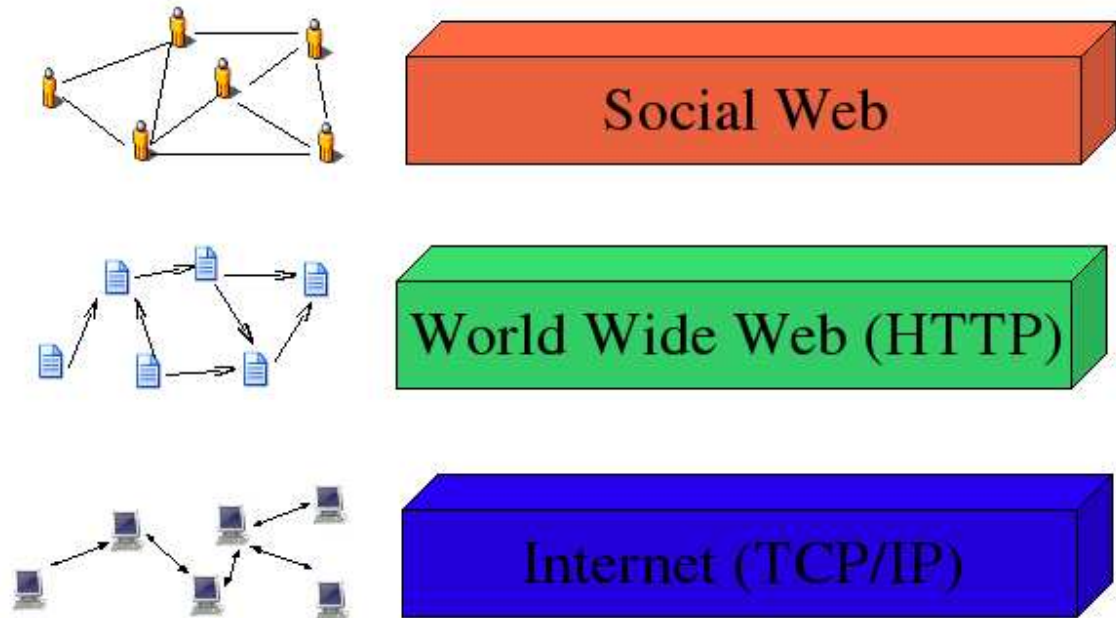
FIGURE 1.2: The Social Web framework. Reproduced from [106]

It has already been suggested to build a social network layer, which may be called "Social Web" on top of the existing HTTP and TCP/IP protocols, as indicated on Figure 1.2. The rationales of building a Social Web are mutual trust and accountability[72]. On the social network where connections are drawn from people's interactions in the real world, participants will be held more accountable for their online activities than they have been in the past. Users can collectively hold those with whom they interact online accountable for their antisocial actions (and for their failures to hold others accountable)[70]. As a result, it can solve or mitigate many problems that plague today's Internet, such as spams, spyware and security issues.

On the other hand, a trustworthy social network facilitates the dissemination of knowledge and information, thanks to the power of word-of-mouth[26][73]. Numerous studies have shown that one of the most effective channels for dissemination of information and knowledge within an organisation is its informal network of collaborators, colleagues and friends[55]. Social networks can spread information much faster than the Web infrastructure, where a website may only be reached if it maintains a high ranking and

visibility on search engines and/or Web portals. Some sources indicated that photo sharing on Facebook was more popular than Flickr[6]. The South Korea-based social network site, Cyworld, which provides blogging, music and video sharing, claimed to have more traffics than the highly touted YouTube[7]. Due to the power of social networks on viral marketing, some singers and artists prefer to promote their music albums on MySpace. As a result, other traditional software such as blogs and Internet forums attempt to exploit the success of SNSs by adding a social networking features.

1.2.4 MAS with Social Intelligence

Social network site can be considered as a large-scaled system of interacting users. Major social networks can have more than 1 million users. The users are actively interacting with each other. Here, the ability to understand human beings and act wisely in human relations is called social intelligence[117], which is equivalent to interpersonal intelligence, a major category in Gardner's classification of multiple intelligences[48]. Social intelligence is crucial to the development of human beings' intelligence. The British psychologist Nicholas Humphrey argued that "the social primates are required by the very nature of the system they create and maintain to be calculating beings: they must be able to calculate the consequences of their own behaviour, to calculate the likely behaviour of others, to calculate the balance of advantage and loss and all this in a context where the evidence on which their calculations are based is ephemeral, ambiguous and liable to change, not least as a consequence of their own actions"[65]. He further suggested that "the social intelligence, developed initially to cope with local problems of inter-personal relationships, has in time found expression in the institutional creations of the savage mind, which has created the systems of philosophical and scientific thought". The novelty of SNSs in the context of MAS is the integration of users' preferences and social networks which exposes the information about the relationships between different intelligent agents that are being guided and supervised by their users, human beings. Given these heuristic information, agents are aware of the relationships of other agents that they interact with. It can be seen which agents are more popular, which are the hubs and which have closer social relationships with whom. The social intelligence will improve the trust, communication and coordination of the agents in a multi-agent system. The social agents have close interactions with human beings. This is particularly true for SNSs that provide development platform and APIs for third party developers, such as MySpace and Facebook. Users can build up their canvas-like software agents by adding various applications, making them full-fledged social agents that are capable to do many jobs and interact with other agents. The development of so-called friendly AI technologies can greatly advance on SNSs. We believe that modelling the connections between human beings can provide an alternative approach to the AI mainstream

---

[6]Facebook Blog: http://blog.facebook.com/blog.php?post=2406207130
[7]Cyworld News: http://www.usnews.com/usnews/biztech/articles/061109/9webstars.cyworld.htm

research methodologies that mainly draw inspirations from modelling a single human being.

1.3   Thesis Structure

The remaining chapters are arranged as follows: Chapter 2 presents a literature review of the history of social network sites in the IT industry, as well as previous research about social science, complex networks and recent research on SNSs. These research and studies provide a historic framework and evidential materials for our research on social network sites. In Chapter 3, we will discuss the issues and problems in social network services. We present the arguments from both publicity and privacy on SNSs. The clash between publicity and privacy, triggered by the use of *static links*, causes friendship inflation, which is a major issue plague today's SNSs. The hyperfriendship network model is introduced in Chapter 4 to analyse the problems incurred by friendship inflation. The theory of friendship inflation is supported by previous research as well as our nearly three-year observation on Facebook users in the network of the University of Southampton. We argue that friendship inflation is responsible for the boom and bust of SNSs. Chapter 5 illustrates the algorithm of *ActiveLink*, which is designed to identify meaningful online social connections. Chapter 6 shows the experimentation and evaluation of key ideas of the algorithm of *ActiveLink*, including *preferential attachment* and *assortativity*. In Chapter 7, we present RealSpace, an SNS model that is based on active links. Finally, in Chapter 8, we will look at the problems in future research.

Chapter 2

Reviews of Social Network

2.1 Introduction

In this chapter, we will review the development of social network sites in the IT industry. This includes the debut of *Sixdegrees.com*, the success of Friendster and now the social network boom. It will be followed by literature reviews from academic accounts, ranging from social science to complex network. Social networks have long been an important research theme in social science. Topics in social science that are particularly relevant to social network research in our thesis are small world phenomenon, strength of weak ties, Dunbar's number, social network analysis and social capital. As data of large-scaled network are increasingly available, social networks are gradually identified as a type of complex network due to its non-trivial topological features. Several mathematical models have been identified to study complex networks. In addition, we present recent research on SNSs. Theses include research on online social capital, privacy issues, friendship performance, impression management, network structure analysis, reputation and trust and so forth.

2.2 History of Social Network Sites

Social network sites attract much attention but is is nowhere a novel idea. The prehistoric social network sites can be dated back to 1995, two years after the momentous Web browser Mosaic was released. *Match.com* was an early online dating site of such type. The website maintained the contacts and profiles of the members which other users could search for. However, this is not the social networking model that is currently being used today. The site did not allow people to interact with one another and share information directly on the site. Users had to communicate with other members either by using email or other offline methods. These sites may be better called "community

8

sites". Examples include *Student.com* and *Classmates.com*. The development of social network sites reflect people's efforts to connect with each other through the Web. Since the launch of *Sixdegrees.com*, we have witnessed a massive growth of social networks. Figure 2.2 illustrates the launch dates of major SNSs.
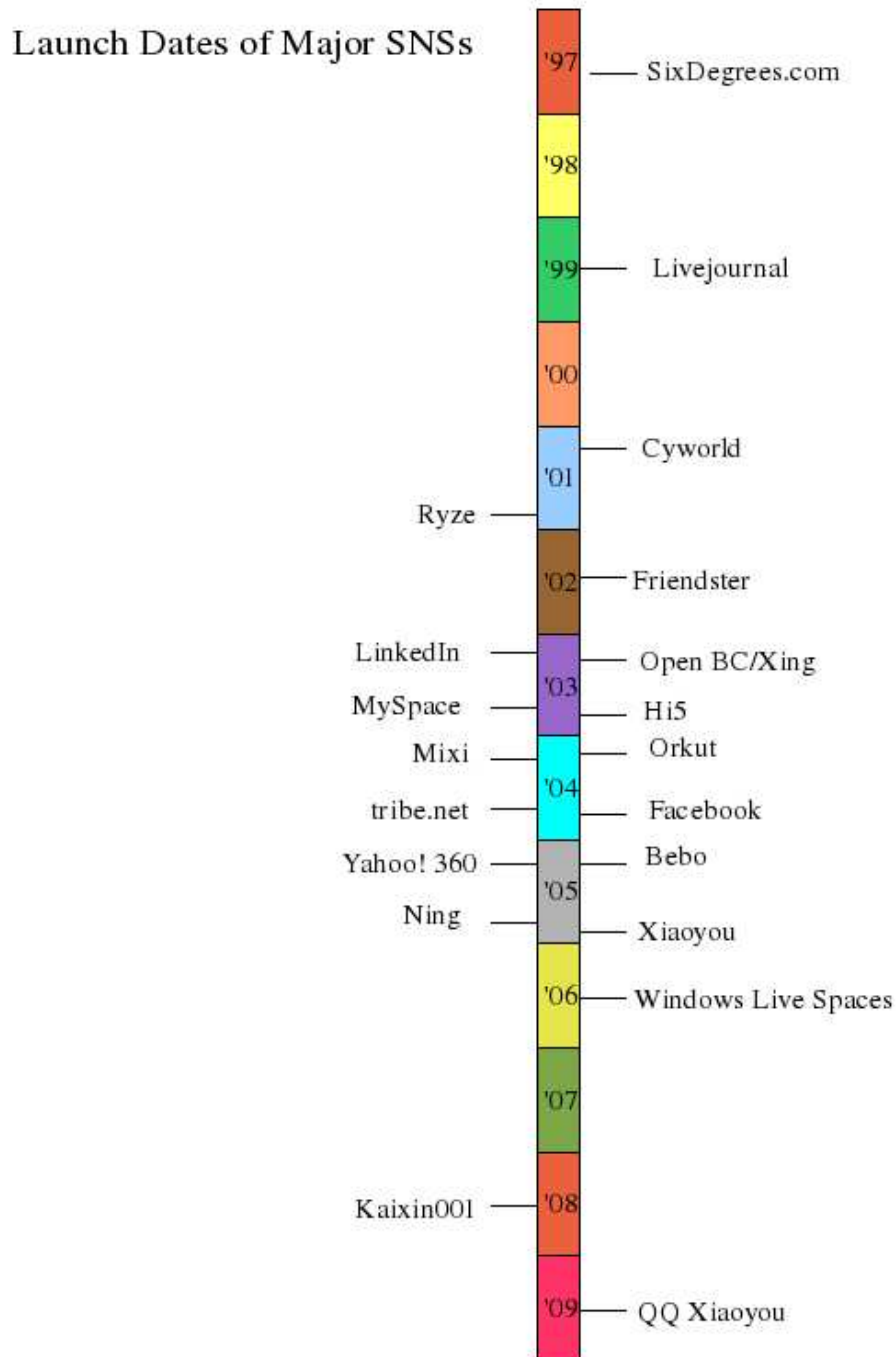
FIGURE 2.1: Timeline of the launch dates of many major SNSs and dates when community sites re-launched with SNS features. Reproduced from [19]

2.2.1 1997-2000: The Debut of SNS

Inspired by the social theory of *six degree of separation*, the site *Sixdegrees.com* was created in 1997. It was the first recognisable social network site[19], as shown in Figure 2.2. Individuals became members by registering information about themselves and listing the names and email addresses of individuals whom they wanted to include as their online friends. SixDegrees promoted itself as a tool to help people connect with their friends[19]. However, the people search functionality was fairly primitive. Members could only be queried by names and emails without much more details about their personal profiles. The site did not provide services such as blogging and photo sharing, which are integral parts of today's SNSs. This is probably because the relevant Web technologies were not yet available or mature at that time. In summary, SixDegrees was simply a bare social network site without much meaningful applications and services. As a result, users often complained that there was little to do on the site after registration.

After nearly four years of operation, the first social network site, with more than 3 million members, closed down at the end of 2000. Figure 2.2 shows the announcement from SixDegrees about closing down its website. Several factors may lead to the site's closure, among those are lack of funding due to the dot-com bubble and unavailability to establish a successful online advertising model. In fact, even today's SNSs are still struggling to find the appropriate business models for social networks. Another noticeable factor is that the Web technology at that time was not mature enough to support meaningful interactions on the site. There is a lack of basic activities such as wall posting, blogging and photo commenting. As commented by the founder Andrew Weinreich, "the site was simply ahead of its time[19]".

However, the website showed a successful demo of *small world phenomenon*. With more than 3 million users, it demonstrated how people can connect with each other effectively on the Web. It justified the encouraging theory of *Six Degree of Separation*. This inspires further development and improvement of SNSs in the following years.

2.2.2 2001-2003: The Success of Friendster

Despite the closedown of *SixDegrees.com* and the dot-com recession in early 2000s, there remain strong interests in developing social network sites. Among those SNSs launched in 2001 were Cyworld and Ryze. Cyworld targeted the South Korea market and has become the biggest social network site in the country. Ryze was designed to target business professionals, particularly new entrepreneurs. Influenced by the success of Ryze, Jonathan Abrams started Friendster in 2002, as a social complement to Ryze[19], competing against *Match.com*, the much more established online dating site. In registration with the site, users were required to create a Friendster profile with answers to questions about personal information such as age, occupation, marital status, general interests,

FIGURE 2.2: *SixDegrees.com* Closing Down

music, books, films and television shows. However, unlike most dating sites of the day, which generally introduced strangers to users, Friendster was seeking to introduce friends of friends to users. A user can invite friends to join his or her personal network and they can write "testimonials" about their friends. The testimonials are often publicly displayed, which often enhance the trust of interaction. Users can navigate the social network within four degrees of their personal network.

The approach is essentially the method that we use on most of the contemporary SNSs. The Friendster social network was then still very primitive but the "four degree of friend chains" had already gave Friendster a huge competitive advantage to its rivals at that time. The site registered its first million users in just six months[105]. It then rocked to 3 million registered members, compared with 85,000 for LinkedIn and Tribe combined[95]. Friendster caught up with the 1995-launched site, *Match.com* in a short period, as indicated in Figure 2.3.

FIGURE 2.3: Friendster's traffic rank (red line) approaches that of Match.com (blue line), according to Alexa Web Traffic

The huge success of Friendster attracted massive press coverage[1], as well as academic research interests. It was reported that Google offered US$30 million to buy Friendster in 2003, but was turned down by Friendster[2]. By January 2004, the site had amassed over 5 million registered accounts and was still growing[16].

2.2.3 2003-Now: The SNS Boom

By 2003 it became obvious that there were huge business potentials of social network sites. Venture capitals were pouring into the SNS industry. MySpace was launched in 2003 to compete with sites like Friendster and Xanga. Some users who were fed up with Friendster were encouraged to join MySpace, with one notable group indie-rock bands[19]. MySpace welcomed the new users and gradually established its reputation as a social network for musicians and their fans. MySpace was taken over in July 2005 for US$580 million by News Corporation, which became the hotspot of the media. Facebook started in a niche market in 2004, catering to the university students in the US. New members were required to register with only university email addresses. In 2005,

---

[1]http://www.jabrams.com/friendster/
[2]http://www.techcrunch.com/2006/10/15/the-friendster-tell-all-story/

FIGURE 2.4: Snapshots of some major Chinese social network sites which are inspired by the success of Facebook

Facebook attracted US$12.7 investment and expanded to include high school students, professionals and finally the general public. Bebo was founded by a British computer programmer in the US and was initially designed for the teenagers. It was acquired by AOL in 2008 for US$850m.

The major Internet players in the industry came to embrace and adopt SNS due to its huge popularity and commercial success. Google launched Orkut in 2004[113]. Yahoo! 360 was developed in 2005. Microsoft renewed its social network platform, Windows Live Spaces. In China, QQ, the most popular instant messaging service in the country, launched social blogging system, QZone, in 2005. It is similar to Microsoft's live spaces. In Japan, Mixi, one of the several SNSs in the country, had over 10 million by May 2008[3]. With many more smaller websites adopting social network technologies, the SNSs keep growing in a fast pace. Figure 2.3 shows the snapshots of some major Chinese social network sites which are inspired by the success of Facebook.

---

[3]http://www.redorbit.com/news/technology/1394023/facebook_still_wants_to_avoid_getting_snatched_up_dealtalk/

2.3 Social Science

Recent advancement in social network sites may be a new phenomenon in the IT industry but social networks have been studied extensively in social science for centuries. Social networks have long been an important research topic in social science. People's connections and the relationships between individuals and groups are key research topics in the study of human beings' society. It is a multi-disciplined area which is contributed by many fields including anthropology, psychology and sociology. Topics in social science that are particularly relevant to social network research in our thesis are small world phenomenon, strength of weak ties, Dunbar's number, social network analysis and social capital. The disciplines provide different perspectives from social computing on social network research as they focus on individual behaviour, institutional incentives and cultural norms[121].

2.3.1 Small World Phenomenon

One of the early studies in social science that generates substantial interests in the research of social network is small world experiment, which was carried out by Stanley Milgram in 1967. They asked 296 people in Nebraska and Boston to pass a letter through acquaintances to a Boston stockbroker. In the end, some 64 letters passed from person to person were able to reach the designated targeted individual. Of those letter chains that were complete, the average number of degrees of separation was 6.2[92]. It should be noted that the people who received the tasks were chosen by random and they passed the letters to people whom they thought would reach the target according to their best knowledge. The idea was later popularised by John Guare in his play *Six Degree of Separation*[58] in 1990. The theory can be illustrated in Figure 2.5.

Inspired by the experiment and the play, a site called *SixDegrees.com* was founded in 1996, with a goal to "find the people you want to know through the people you already know"[22]. SixDegrees was one of the early successful websites based on the theory of *Six Degree of Separation*. Watts et al. attempted to explain the small-world phenomenon by developing a statistical model, which was published on the *Nature* Journal in 1998 and attracted a lot of research interests. To further examine the small world theory in a more rigorous way, Dodds, Muhamad and Watts conducted a research on global social search in 2003 and found that social searches can reach their targets in a median of five or seven steps[31]. In their experiment, more than 60,000 email users attempted to reach one of 18 target persons in 13 countries by forwarding messages to acquaintances. They find that successful social search is conducted primarily through intermediate to weak strength ties, does not require highly connected "hubs" to succeed, and, in contrast to unsuccessful social search, disproportionately relies on professional relationships. Interestingly,
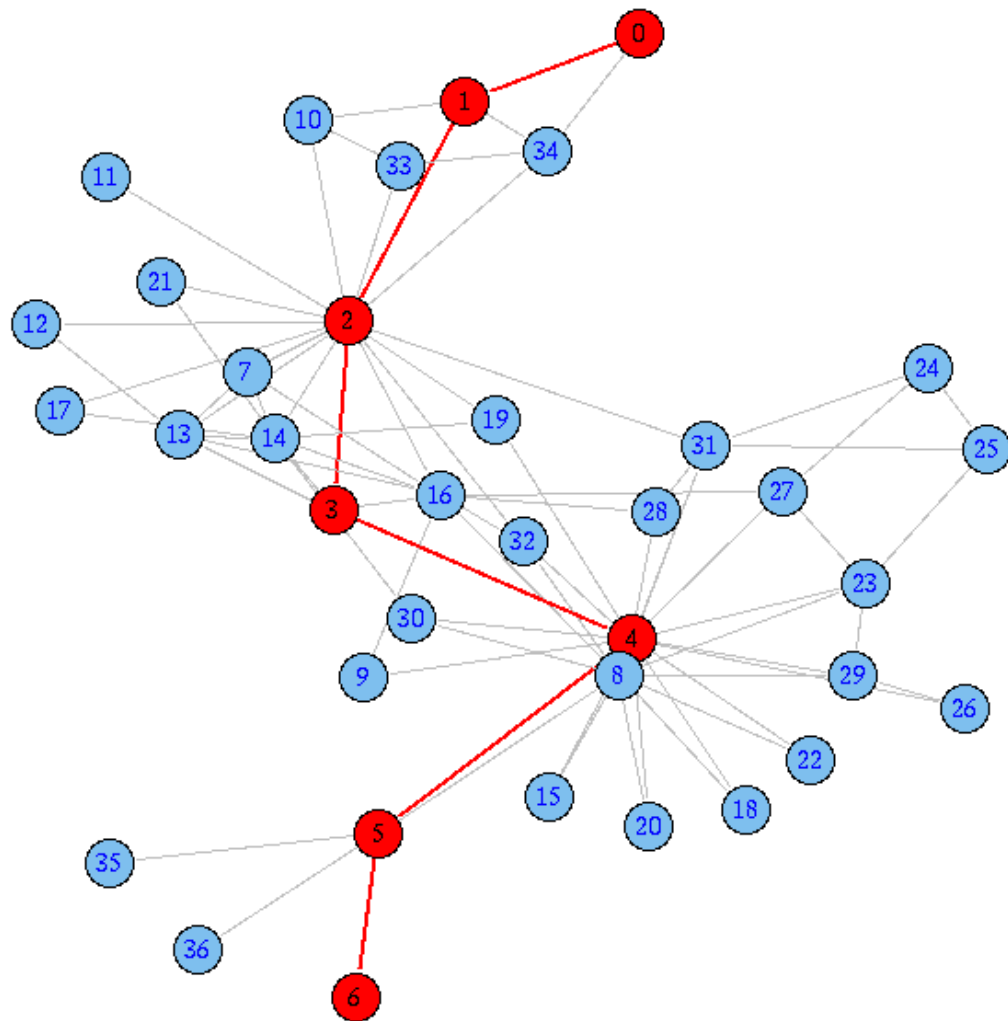
FIGURE 2.5: Illustration of the Six Degrees of Separation theory

they discovered that although global social networks are, in principle, searchable, actual success depends sensitively on individual incentives. A newer research conducted by Microsoft researchers on its instant-messaging system shows that the average path length is 6.6[85]. The data was taken from the MSN conversations during June 2006. The dataset contains 180 million nodes and 1.3 billion undirected edges. Noted that unlike a buddy graph where two people are connected if they appear on each other's contact lists, this so called communication network is from the data where each user is represented by a node and an edge is placed between users if they exchanged at least one message during the month of observation.

2.3.2   Strength of Weak Ties

The ties of social network refer to the connections between people and organisations. There are three categories of ties: strong, weak, and absent. It is naturally assumed that strong ties are important in one's social network as they directly link to the user per se. On the contrary, weak ties were often considered as less important if not ignored by the researchers. However, a paper titled *The Strength of the Weak Ties* was published by Mark Granoveter in 1973, in which he argued that weak ties are crucial in interpersonal networks as they enable reaching populations and audiences that are not accessible via strong ties[55]. The somewhat counter-intuitive discovery is considered one of the most influential sociology papers ever written[9]. The subsequent research show that the weak ties may be explained by the diffusion of ideas and innovation[56]. Interestingly, Bernd Wegener suggested that individuals with high status prior jobs benefit from weak social ties, whereas individuals with low status prior jobs do not[124]. Burt argued that structural holes, a concept closely related to weak ties, are entrepreneurial opportunities for information access, timing, referrals and control[23]. With the emerge of social network software, where people can make new connections by a few clicks, it is much more convenient to establish weak ties.

2.3.3   Dunbar's Number

In the early 1990s, based on the research on non-human primates, the British anthropologist Robin Dunbar theorised that there is a supposed cognitive limit to the number of individuals with whom any one person can maintain stable social relationships[37]. He proposed that the number is approximately 150, which is now known as Dunbar's number. He argued that the limit exists not only in ancient villages and tribal groups but also in modern armies and companies. Gladwell elaborated the idea and proposed the rule of 150[50]. He suggested that groups of less than 150 members usually display a level of intimacy, interdependency and efficiency that begins to dissipate markedly as soon as the group's size increases over 150.

Recent research indicates that an upper limit may indeed exist on SNSs. Golder et al. found that "(on Facebook) thousands upon thousands of people have anywhere from 1 to a few hundred friends, but at about 250 friends, the number starts to drop sharply[53]". Some SNSs recognise such limit and try to reduce the number of friends. Facebook, for instance, was reported to impose a 5000-friend limit on users with "excessive friends". Given the rapid increase of social network connections, questions are raised on how to maintain meaningful relationships[118].

2.3.4 Social Network Analysis

Social scientists have established a set of techniques for analysing social networks. It is called "social network analysis", or SNA, which is focused on identifying patterning of people's interaction[120][110]. It is based on the assumption that social structure can affect people's behaviours and activities. Social network analysis is essentially a variant of network analysis, which is a branch of graph theory. SNA plays a crucial role in sociology, anthropology, social psychology, organisational studies and business. Measures in social network analysis include betweenness, closeness, degree centrality, eigenvector centrality and so on. The term **centrality** refers to the relative importance of a node within the network. The techniques are fairly established and have been applied to many areas. For example, it was used to analyse and track down the terrorist groups[78]. In computer science, it is an important technique in knowledge management, ubiquitous computing, information retrieval and so forth. Google's ranking algorithm, PageRank, for example, is a variation of eigenvector centrality over Web pages[21].

2.3.5 Social Capital

Social capital is an "umbrella concept". It is an elastic term with a broad range of definitions[3]. Among the early definitions, social capital is defined as the aggregate of the actual or potential resources which are linked to the possession of a durable network of more or less institutionalised relationships of mutual acquaintance and recognition – or in other words, to membership in a group – which provides each of its members with the backing of the collectivity-owned capital, a "credential" which entitles them to credit, in the various senses of the word[15]. Others, like Coleman, stated that social capital broadly refers to the resources accumulated through the relationships among people[27]. Given the ambiguity of the concept, it is not easy to provide an accurate measure of social capital. It may be measured by resources in the social network. It may also be measured as an outcome of the network effect rather than the network itself[125].

A growing body of literature has confirmed that social capital is correlated with positive individual and collective outcomes in areas such as better health, lower crime, better educational outcomes, economic development and good government[108][127]. Social capital, as measured by the strength of family, neighbourhood, religious and community ties, is found to support both physical health and subjective well-being[60]. Social capital researchers have found that various forms of social capital, including ties with family, friends and neighbours, are associated with positive psychological and social outcomes[11].

2.4 Complex Network

As data of large-scaled networks are increasingly available, social networks are gradually identified as a type of complex network due to its non-trivial topological features that are not present in simple networks such as random graphs and lattices. In fact, it is a major category of complex networks, which also include but are not restricted to computer networks, biological networks and transport networks[97]. Complex network normally exhibits robust organising principles. The study of complex network may be regarded as an intersection between graph theory and probability theory. In the past few years, the advance of information technology led to the emergence of large databases on the topology of various social networks. Computing powers allowed researchers to investigate the statistical properties of networks containing millions of nodes, exploring questions that could not be addressed previously. Many new concepts and measures have been proposed and investigated in depth in the last decades. The study on various data of complex network supports the idea that the networks have three robust measures of topology[99][5][97][36][29][28]: small average path length between any two nodes (small-world effect), presence of cliques or large clustering coefficient, and power law degree distribution (scale-free). Some underlying principles have been identified for explaining these topological characteristics. For instance, short paths could provide high-speed communication channels between distant parts of the system, thereby facilitating any dynamical process that requires global coordination and information flow[114]. Large clustering coefficient means that on average a person's friends are far more likely to know each other than two people chosen at random[121]. It is also known as transitivity in sociology[120]. In particular, transitivity, which is derived from *balance theory*, has been proposed as a fundamental social law[120]. For power-law degree distribution, Albert, Jeong and Barabasi suggested that scale-free networks are resistant to random failures because a few hubs dominate their topology[6]. The existence and persistence of these interesting characters in social network as well as other complex networks inspired researchers to look for new mathematical models for network study.

2.4.1 The Erdos-Renyi Model

In their classic article on random graphs, Erdos and Renyi proposed a simple model of a random network. Take some number of n nodes and connect each pair with probability p. This defines $G_{n,p}$ in the ER model[44]. Figure 2.6 illustrates the graph evolution process for the ER model. Given the limit of large n, the mean degree z is $p(n-1)$, in which case the model has a Poisson distribution. The typical distance through the network is $l = \lg g / \lg z$, which shows a relatively short average path. The model is well known for the study of connectedness of random graph, however, the model fails to describe other significant features such as clustering and degree distribution that also exist in real-world social network.
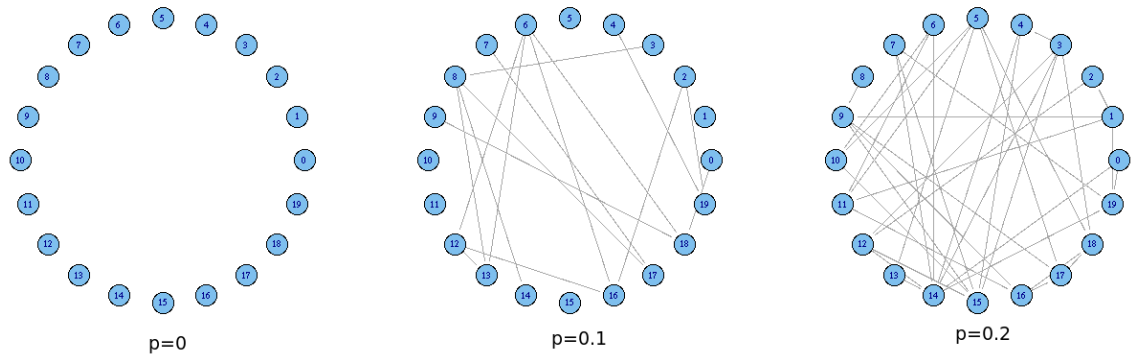
FIGURE 2.6: Illustration of the graph evolution process for the ER model. Reproduced from [5]

2.4.2 The Watts-Strogatz Model

The real-world social networks are well connected and have a short average path like random graphs, but they also have exceptionally large clustering coefficient, which had not been captured by ER model and other random graph models. Watts and Strogatz proposed a one-parameter model that interpolates between an ordered finite dimensional lattice and a random graph. The algorithm of the model is shown as follows (Figure 2.7): Starting from a ring lattice with n vertices and k edges per vertex, each edge is rewired at random with probability p[123]. Watts et al. found that $L \sim n/2k \geq 1$ and $C \sim 3/4$ as $p \to 0$, while $L = L_{random} \ ln(n)/ln(k)$ and $C = C_{random} \ k/n \leq 1$ as $p \to 1$. The clustering coefficient has been much investigated for the model. It concludes that the WS network is suitable for explaining such properties in many real-world examples.
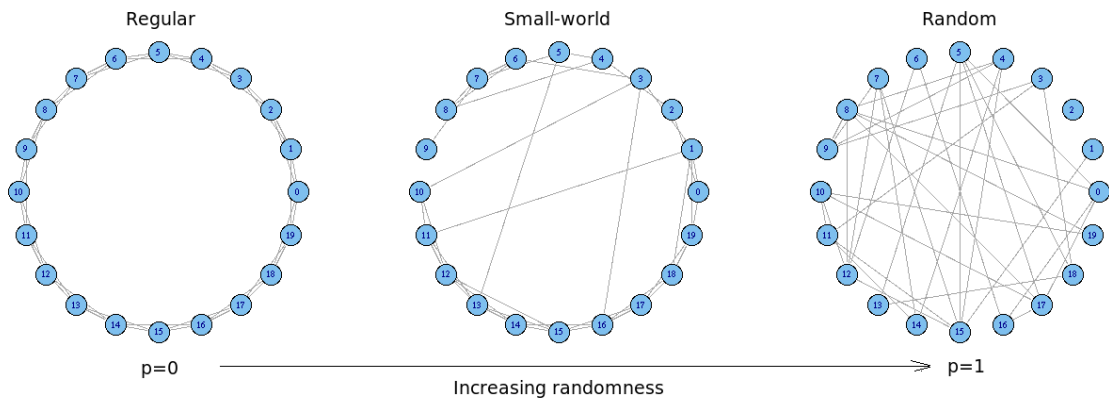
FIGURE 2.7: The random rewiring procedure of the WS model which interpolates between a regular ring lattice and a random network. Reproduced from [123]

The model has been studied widely since the details have been published. It is particularly important in the study of the small-world theory. Some important search theories such as Kleinberg's work is based on a variant of the model. The disadvantage of the

model, however, is that it has not been able to capture the power law degree distribution as presented in most real-world social networks.

2.4.3 The Barabasi-Albert Model

The previous two models take observed properties of real-world networks and attempt to create models that incorporate those properties. However, they do not assist to understand the origin of social networks and how they generate those properties as they evolve. Barabasi and Albert proposed a model that tried to address these problems. There are two important hypotheses with the model[10]:
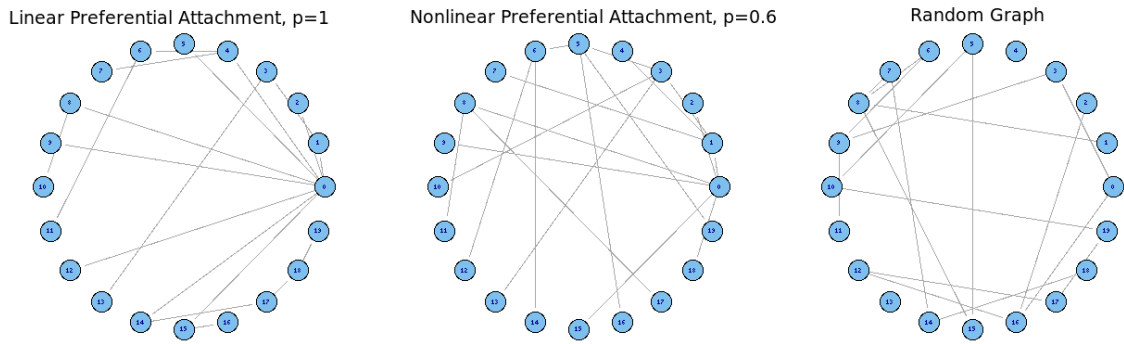
FIGURE 2.8: The graph generated from BA model has more hubs than that generated from ER model.

(1) *Growth*: Let p_k be the fraction of nodes in the undirected network of size n with degree k, so that $\sum_k p_k = 1$ and therefore the mean degree m of the network is $\frac{1}{2} \sum_k k p_k$. Starting with a small number of nodes, at every time step, we add a new node with m edges that link the new node to old nodes already presented in the system.

(2) *Preferential attachment*: When choosing the nodes to which the new node connects, the probability that a new node will be connected to a node of degree k is:

$$\prod = \frac{k p_k}{\sum_k k p_k} = \frac{k p_k}{2m} \tag{2.1}$$

Using master-equation approach, it can be shown that:

$$p_k = \begin{cases} \frac{2m(m+1)}{(k+2)(k+1)k} & for \quad k > m \\ \frac{2}{m+2} & for \quad k = m \end{cases} \tag{2.2}$$

It has been pointed out that the concept of *preferential attachment* is largely influenced by the notion of *cumulative advantage* in Price's model[97]. In the limit of large k it gives a power law degree distribution $p_k \sim k^{-\alpha}$, with the $\alpha = 3$. Figure 2.9 shows the
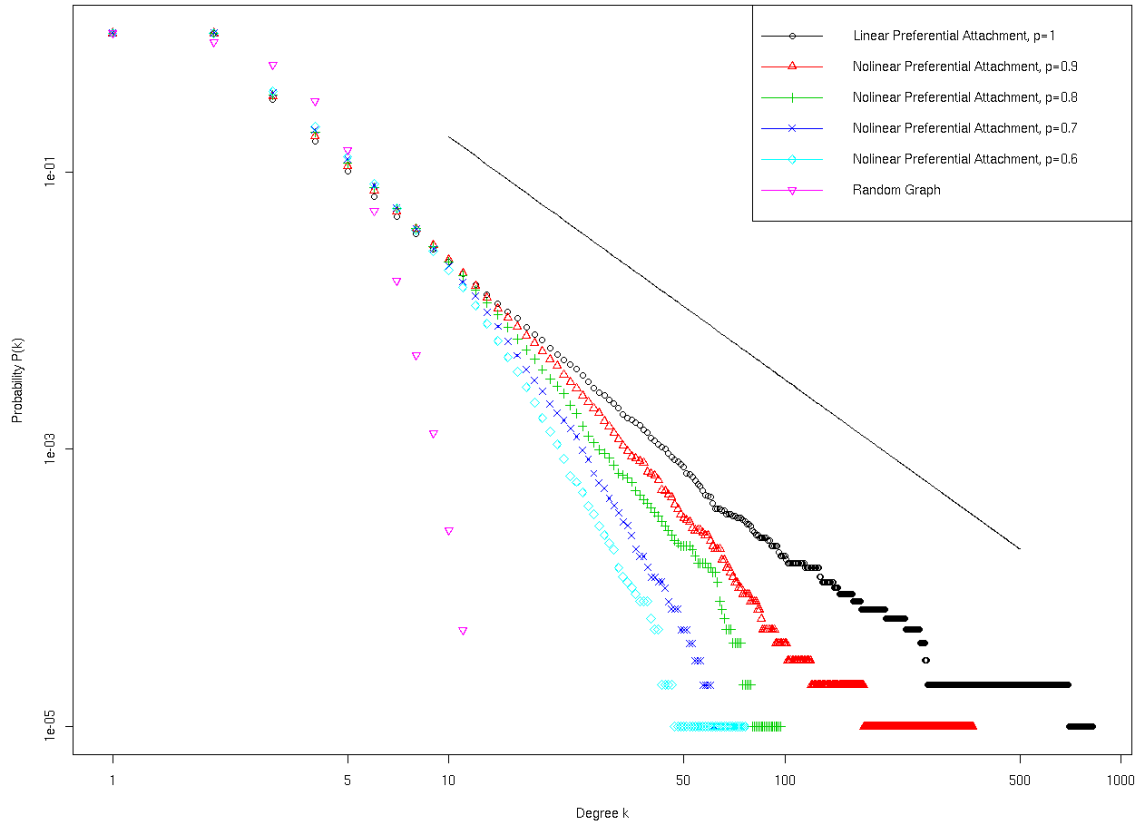
FIGURE 2.9: Degree Distribution for BA model with different exponents of the preferential attachment process.

degree distribution for the model. While the BA model captures the power law tail of the degree distribution, it has other properties that may or may not agree with empirical results in real networks. Recent analytical research on average path length indicate that $l \sim ln(N)/lnln(N)$. Thus the model has much shorter l than that of a random graph. The clustering coefficient decreases with the network size, following approximately a power law $C \sim N^{-0.75}$. Though greater than those of random graphs, it is dependent on network size, which is not true for real-world social networks.

Two limiting cases have been developed to test the two hypotheses of the model. Model A keeps the growing character of the network without *preferential attachment*. Barabasi *et al.* found that p_k decays exponentially, indicating that the absence of *preferential attachment* eliminates the scale-free character of the resulting network. Model B removes the growth process whilst maintaining the *preferential attachment*. Through numerical simulations, they found that while at early times the model exhibits power-law scaling behaviour, p_k is not stationary and it eventually becomes nearly Gaussian around its mean value. The failure of models A and B to lead to scale-free distribution indicates that both *growth* and *preferential attachment* are needed simultaneously to reproduce the stationary power-law distribution observed in real networks.

2.4.4 Community Structure

Community structures are groups of nodes which are more densely interconnected with each other than with the rest of the network. This can be easily seen in social network. It is common experience that people do divide into groups along lines of interest, occupation, age, and so forth[97]. It is therefore widely assumed that community structure is one of the characteristics of social networks[120][110]. Thus, it is of great benefit to identify the community structure in large-scaled networks where network structures are not easy to perceive.

The traditional method for detecting community structure is hierarchical clustering[49], as shown in Figure 2.10. However, the method fails to detect the peripheral vertices. Another algorithm that was developed recently is "edge betweenness", which is the count of how many geodesic paths between vertices running along each edge in the network[97]. Newman and Girvan[98] developed an algorithm based on *modularity* to overcome the problem.
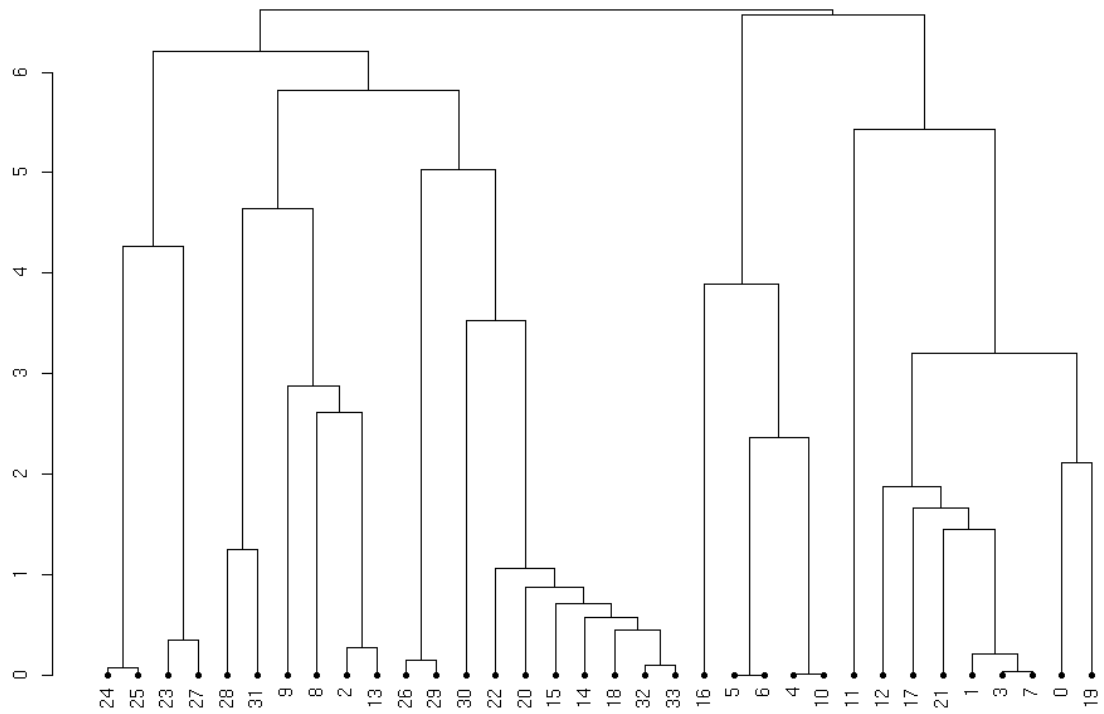
FIGURE 2.10: Illustration of Hierarchical Clustering Algorithm for Community Structure Detection

2.4.5 Searching in Social Network

The major objective of study on the structure of networks is to understand and explain the functioning of the systems built upon the networks. Important dynamical processes taking place on social networks include epidemiological processes, spreading of ideas,

innovations and computer viruses, diffusion innovation, and information searching. The network topology usually plays a crucial role in determining the system's dynamical features. In this section we review some important models and theories on network searching.

Kleinberg's Lattice Network

Kleinberg proposed a model based on WS model to explain that why arbitrary pairs of strangers be able to find short chains of acquaintances that link them together[76]. The model employs a two-dimensional lattice (with size $n \times n$) as basic structure. Notice that it was NOT a ring structure as originally proposed in WS model. Whilst all the nodes in the ring model have the same number of connections, the nodes in the out-most area of the lattice structure will have less connections than others due to the grid structure. Each node has a directed edge to every other node within lattice distance p – these are its *local contacts*. p is very small, meaning each node only knows its neighbours for some number of steps in all directions. On the other hand, the node has directed edges to q other nodes, $q \geq 0$. Each number of acquaintances distributed across the grid. Figure 2.11 shows the graph of the lattice.
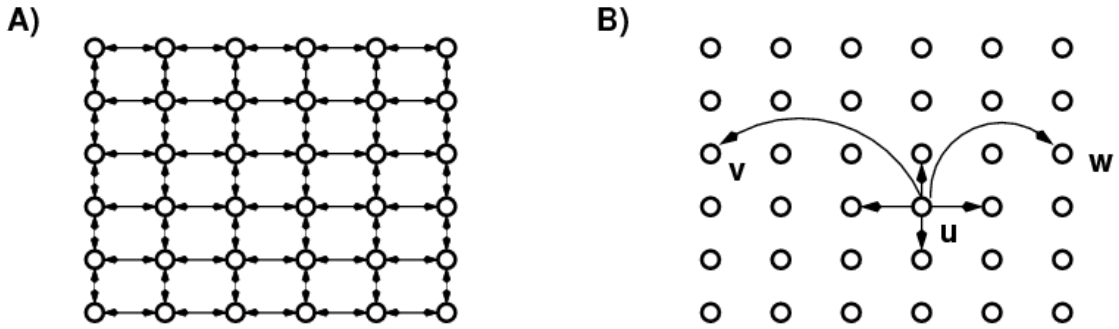
FIGURE 2.11: (A) A two-dimensional grid network with n=6, p=1, and q=0; (B) p=1 and q=2, v and w are the two long-range contacts. Reproduced from [76]

The probability that such edge exists is

$$d^{-r} \tag{2.3}$$

Here $r \geq 0$ and d is the lattice distance between the node and its remote acquaintance, also known as *long-range contact*. Kleinberg proved the following statements:

(a) For $0 \leq r < 2$, there is a constant c, depending on p, q, r, but independent of n, so that the expected delivery time of any decentralised algorithm is at least $cn^{(2-r)/3}$.

(b) For $r = 2$, there is a constant c, depending on p, q, r, but independent of n, so that when p=q=1 the expected delivery time of any decentralised algorithm is at most $O(logn)^2$.

(c) For $r > 2$, there is a constant c, depending on p, q, r, but independent of of n, so that the expected delivery time of any decentralised algorithm is at least $cn^{(r-2)/(r-1)}$.

The decentralised algorithm achieving the bound in (b) is as follows: each node forwards the message to a neighbour — *long-range* or local — whose grid distance to the target is as small as possible. This is in fact a simple greedy algorithm in which at each step along the way the message is passed to the person that the current holder believes to be closest to the target.

The proof has been demonstrated to be true on hierarchical models and partially applied to set systems[77]. Kleinberg's proof reveals an important feature of search in social networks: the existence of short paths lies not on the sophistication of the search algorithms but on the topological structure of the network. As long as the networks have topological characters shown in WS model, there can always be short paths between any two nodes and the paths can be constructed by message carriers with only local knowledge.

Search on "Social Distance"

Kleinberg's model indicates that one needs not worry about the greedy algorithm performed by individual but should rather focus on the whole network topology. It does not, however, give a thorough investigation of how such uncoordinated search behaves.

Empirical experiments carried out by sociologists show that people navigate social networks by looking for common features and similarities between their friends and the targeted individuals[75]. They pointed out that the top choices for choosing a friend are location and occupation. Watts *et al.* proposed a model for a social network that is based on social grouping[122]. There are two major settings with the model:

(1) Individuals belong to groups which in turn belong to groups of groups and so on giving rise to a hierarchical categorisation scheme, as shown in Figure 2.12.
(2) The model have many hierarchies indexed by $h = 1...H$. These H dimensions of hierarchies are independent of each other. The social distance between any two nodes takes the minimum ultrametric distance over all hierarchies.

The search algorithm allowed the individuals to have two kinds of information: social distance, which can be measured globally but is not a true distance; network paths, which generate the true distances but are known only locally. They found that such an algorithm performs well over a broad range of parameters. One interesting result is that the best performance is achieved for H=2. They believe the number conforms to the empirical evidence that individuals across different cultures in small-world experiments typically utilise two or three dimensions when forwarding a message.

Kleinberg found in similar model that the search can be completed in $O(logn)$ steps[77]. Based on the result of computer simulation, Simsek and Jensen[111] suggested that a
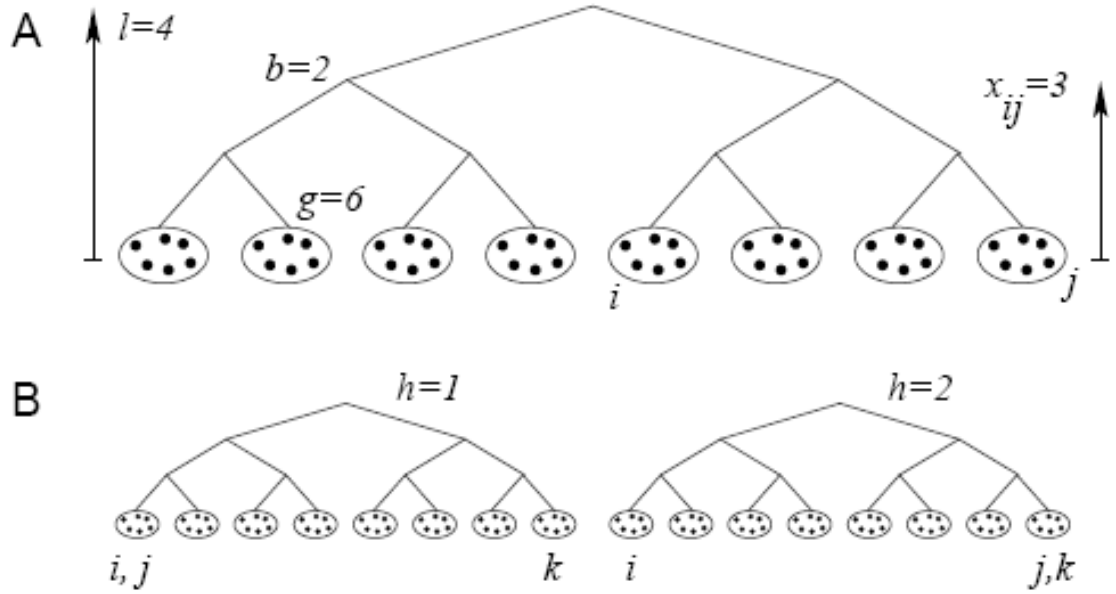
FIGURE 2.12: The Hierarchical "Social Distance" Tree Model. Reproduced from [122]

heuristic decentralised algorithm taking both social distance and node degree information can perform more efficiently than using only one of these factors.

2.5 Recent Research on SNSs

Recent surge of popularity in SNSs makes social networks the hotspot in social computing study. Social computing concerns with the social behaviours and computational systems. Emails, instant messengers and blogs fall into this category of the research. For social network study, the bulk of research concentrate on online social capital, privacy issues, friendship performance, impression management and network structure analysis. We will give a detailed review on these research in this section.

2.5.1 Online Social Capital

Investigations on the motivation and purpose of using social networks have been conducted by researchers. Lampe et al. found that users of Facebook largely employ the site to learn more about people they meet offline, and are less likely to use the site to initiate new connections[79]. The most likely activities are "Keep in touch with an old friend or someone I knew from high school" and "Check out a Facebook profile of someone I met socially" while "Finding casual sex partners" is the lowest in the expectation scale. This mirrors the success of SNSs to encourage the use of genuine identity in social networking. Joinson pointed out seven unique uses and gratifications: social

connection, shared identities, content, social investigation, social network surfing and status updating[71]. Ellison found that there is a robust connection between Facebook usage and indicators of social capital[41]. Analysis suggests that a strong association between the use of Facebook and the three types of social capital (bonding, bridging and maintained capital), with the strongest relationship being *bridging social capital*[42]. Omurchu et al. found that communities can be informed more quickly through online social networking, and become more engaged and involved with one another in an era when social capital in the offline world is on the decline[101].

Some research study the ages on the social networks. Valkenburg, Peter and Schouten found that positive feedback on the profiles enhanced adolescents' social self-esteem and well-being[119]. Wu et al. found that in addition to an individual's own human capital and network position, the human capital and status of her friends can be instrumental to her success[127]. Other research indicates that teens are more accessible and are more likely to make new friends through social network sites[82].

2.5.2 Privacy Issues

In the early days of SNSs, users usually ignore the privacy settings. Gross and Acquisti[57] discovered that limiting privacy preferences are hardly used and only a small number of members changed the default privacy preferences. They argued that social networks such as Facebook and MySpace are different from traditional online communities in that there are public linkages between an individual's profile and the real identity of its owner, as well as the perceived connection to a physical and ostensibly bounded community. Some users manage their privacy concerns by trusting their ability to control the information they provide and the external access to it[1]. However, other research suggests that there are significant misconceptions among some members about the online community's reach and the visibility of their profiles. A social network provider will sometimes violate its terms and conditions about privacy. Rosenblum[109] singled out unauthorised use by third parties as one of the external risks. Stutzman compared SNS with traditional methods for identity information disclosure, such as a campus directory and found that social network fosters a more subjective and holistic disclosure of identity information[115]. Patil and Lai revealed that presenting participants with a detailed list of all pieces of personal context to which the system had access, did not result in more conservative privacy settings[103]. They showed that although location was the most sensitive aspect of awareness, participants were comfortable disclosing room-level location information to their team members at work.

Barnes discussed teens behaviours on social networks and she raised the privacy paradox that on one hand, teenagers reveal their intimate thoughts and behaviours online but, on the other hand, government agencies and marketers are collecting personal data about them[12].

Disclosure of personal information usually attracts spamming and phishing. Zinman and Donath found that it is more difficult to detect spams in SNSs than in emails because unsolicited messages no longer mean unwanted in social network sites[128]. Other work suggests that SNSs identify "circles of friends" that allow a phisher to harvest large amounts of reliable social network information[67]. In a survey of 2,117 Americans with 1,017 of them being Internet users, Fox pointed out that 86% Internet users are in favour of "opt-in" privacy policies that require Internet companies to ask people for permission to use their personal information[47].

2.5.3 Friendship Performance

It takes only a few steps to befriend with one another in social network sites. Any user can send a friend request to one another, who normally accepts the request without a second thought. It takes a few clicks to finish the process. However, the convenience of friending raises the question of the meaning of friendship. Boyd noted that when traversing the network on Friendster, there is no way to determine what metric is or what the role or weight of the relationship is[16].

Donath and Boyd explored the social implications of the public display of one's social network. Social status, political beliefs, musical taste, etc, may be inferred from the company one keeps[34]. Fono and Raynes-Goldie studied the friendship on Livejournal and found that user opinions, behaviours, understandings and attitudes varied widely[46]. Boyd argued that the established friending norms evolved out of a need to resolve the social tensions that emerged due to technological limitations and friending supports pre-existing social norms[17]. She argued that the example of Friendster demonstrates the inverse relationship between the scale of social network and the quality of the connections within the network[18]. Dwyer noted in her survey that participants acknowledged the friendships were "superficial"[38]. After a survey on Facebook usage, Tong et al. raised the doubts about Facebook users' popularity and desirability[118]. They showed that there exists a curvilinear effect of sociometric popularity and social attractiveness emerged, as did a quartic relationship between friend count and perceived extraversion. Golder, Wilkinson and Huberman questioned the problematic status of the "friend" links[53]. They proposed that messaging should be perceived as a more reliable measure of Facebook activity. Laraqui designed a social network system utilising users' activities[80]. Huberman et al. developed a mobile social network application for close relationships[64][7].

The Facebook Data Team has recently published a blog about the analysis of the social relationships on Facebook[4], entitled "Maintained Relationships on Facebook". They found that on Facebook the number of the reciprocal relationships, where reciprocal communications taking place between two parties, is far less than that of the maintained

---

[4]http://www.facebook.com/note.php?note_id=55257228858

relationships, where a user had clicked on another's News Feed story or visited their profiles more than twice, as shown in Figure 2.13. Wilson et al. questioned if the social links of SNSs are valid indicators of real user interaction. They proposed the *interaction graph* as a more accurate representation of meaningful peer connectivity on social networks[126].
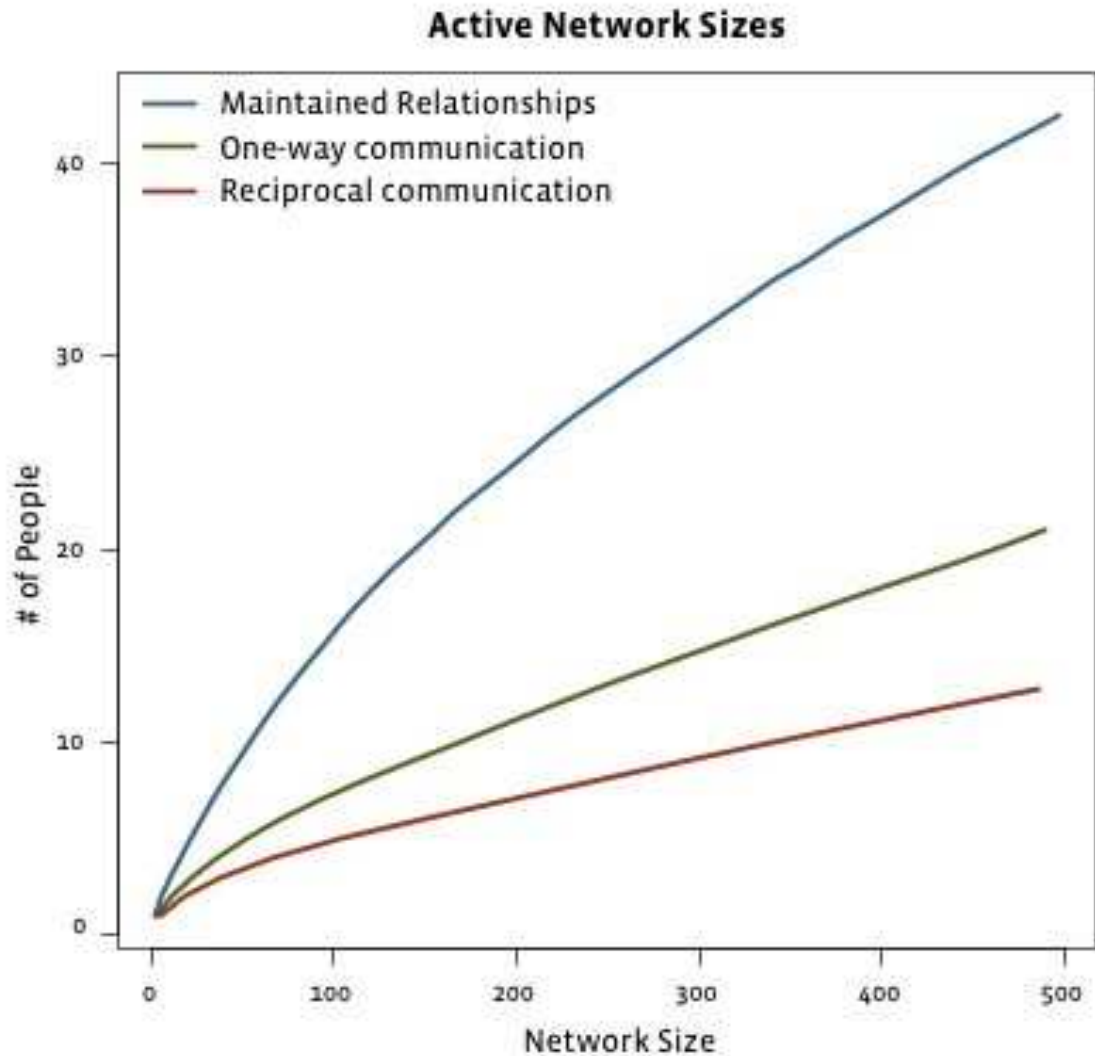
FIGURE 2.13: The number of the reciprocal relationships is far less than that of the maintained relationships on Facebook. Taken from Facebook Blog.

2.5.4 Impression Management

Impression management is a process through which users try to shape the impressions other people form of them. In SNSs, users can control the impressions by manipulating profiles, friend listings and behaviours on the sites. Users' main goal is to keep their totality and coherence. They will adjust themselves to different contexts offered

to them[51]. Swinth, Farnham and Davis described the results of their study which indicated that individuals will provide more personal information when completing profiles for online communities that facilitate deeper and more meaningful interpersonal relationships[116]. Boyd[16] noticed that on Friendster, context is missing from what one is presenting. She also pointed out that most users fear the presence of boss and mother, suggesting that users are aware that in everyday activity they present different information depending on the audience. Markwick argued that current SNSs encourage a commodified, fixed, singular view of identity presentation that limits their usefulness for network mapping and relationship building[89][74]. DiMicco and Millen examined online profile pages and interviewed employees at a large software development company and found that there are difficulties in simultaneously using a single site for both professional and non-professional use[30].

Other works suggested that when expectations created by a profile did not match reality, relationships were severed[38]. Additionally, Boyd and Heer pointed out how the performance of social identity and relationships shifted the profile from being a static representation of self to a communicative body in conversation with the other represented bodies[20].

2.5.5   Network Analysis

As mentioned earlier, social network has long been studied as a type of complex network and significant progress has been made recently[5][97]. Many researchers therefore attempt to check if these theories and models of complex network still hold for online social networks. Having examined the data on *pussokram.com*, Holme et al. observed that there is no apparent cut-off in the degree distribution of the network due to the low cost of acquiring new contacts[63]. They also found that reciprocity is rather low and mixing by degree between vertices is dissortative, which is different from most real-world social networks. Adamic and Adar studied the network of club Nexus and observed the social network phenomena such as the small world effect, clustering, and the strength of weak ties[2]. Liben-Nowell et al. studied the social network of LiveJournal and showed that one third of the friendships are independent of geography[88]. Lee, Kim and Jeong noticed that the quantities related to the properties such as degree and betweenness centrality, distribution and assortativity in sampled networks appear to be quite different for each sampling method[81]. Mislove, Marcon and Gummadi studied Flickr, YouTube, LiveJournal and Orkut[93]. While their results confirmed the power-law, small-world and scale-free properties of online social networks, they also found that the assortativity is different from other previously observed power-law networks. Other work suggested that the network on Slashdot exhibited moderate reciprocity and neural assortativity by degree[54]. Yuta, Ono and Fujiwara investigated the topology of Mixi and found that there exists a range of community-sizes in which only few communities are detected[83]

Ahn et al. compared Cyworld, MySpace and Orkut and they demonstrated that Cyworld data's degree distribution exhibits a multi-scaling behaviour. They conjecture that Cyworld's testimonial network is more similar to off-line social networks than its friends' network[4]. As online activity is an effective means for measuring the dynamics of SNSs. The messaging on Facebook was studied by Golder et al., who discovered the temporal messaging patterns[53].

2.5.6 Reputation and Trust

Social network sites, with their detailed profiles and connections of the members, can build reputation and trust into the existing Web infrastructure[72]. Some issues confronting today's Internet include spams, spyware and security. Social network sites address these issues by establishing peer production of governance[70]. Different models of trust and reputation have been proposed. For example, Huynh et al. discussed interaction trust, role-based trust, witness reputation and certified reputation[66]. Golbeck and Hendler studied the trust relationships in web-based social networks and proposed algorithms for inferring trust relationships between individuals that are not directly connected in the network[52]. Dwyer, Hiltz and Passerini made an online survey of Facebook and MySpace and found that Facebook members expressed significantly greater trust in both Facebook and its members while MySpace members reported significantly more experience using the site to meet new people[39].

2.5.7 Other Research

One of the strengths of social network is its power of viral marketing. Pedro Domingos demonstrated in their experiment that it is possible to achieve much higher profits than if ignoring interactions among customers and the corresponding network effects, as traditional marketing does[32]. The rationales behind this are that on social network, a set of customers such as opinion leaders can be specifically targeted to maximise word of mouth. Leskovec et al. showed a model based on social network that can identify product and pricing categories for which viral marketing seems to be effective[84]. Social network system can be applied to support knowledge sharing between people[26]. Ermecke, Mayrhofer and Wagner surveyed data of 475 Facebook users and found that active (purposeful recommendations form peers) viral channels dominate in convincing users to actually start using a product or service[45].

Murnan compared the use of online social network with email, suggesting that email is no longer the only communication method by students[94]. Nyland observed that more religious individuals are more likely to use social networks to maintain already existing relationships[100]. With an examination on the language use on LiveJournal, Herring et al. revealed that English dominates globally but not locally and network robustness

is determined mostly by population size[62]. Mazer, Murphy and Simonds found in an experiment that participants who accessed the Facebook website of a teacher high in self-disclosure anticipated higher levels of motivation and affective learning and a more positive classroom climate[90]. Charnigo and Barnett-Ellis looked at the impact of online social networks on academic libraries[24]. Some interesting findings suggest that some enthusiastic librarians wanted to use Facebook to promote library services and events. Baron compared the evolving relationship between social network platform and instant messenger platform[13]. Snyder et al. argued that SNSs such as MySpace need to introduce social contract theory to enforce the rules for online activities[112]. Backstrom et al. found that the tendency of an individual to join a community is influenced not just by the number of friends he or she has within the community, but also crucially by how those friends are connected to one another[8].

2.6   Summary

In this chapter we have summarised prior work in the literature which is relevant to our study of social network system. In the following chapter we presented the clash between publicity and privacy, triggered by the use of *static link*. It causes the problem of friendship inflation, which will be analysed in more details in Chapter 4. An algorithm called *ActiveLink* is designed to identify meaningful online social connections and will be discussed in Chapter 5.

Chapter 3

Challenges of Friendship Management

3.1   Introduction

As we discussed earlier in the first chapter, social network sites are defined as websites that allow users to traverse others' social network of more or less institutionalised relationships of mutual recognition. The profiles on social network sites are essentially public display of private profiles, of which only the owners have full knowledge. The relationships shown on the sites are public exhibition of private relationships, of which only the owners and their friends have full knowledge. On one hand, system designers hope that every member can publicise their profiles and friend lists so that the real-world network can be traversed and navigated effectively. The publicity can substantially increase users' online social capital, improve the diffusion of information and knowledge and enhance online trust and security. The publicity of the individuals will in turn benefit every individual. On the other hand, from individuals' point of view, while some publicity will benefit them, too much publicity will always incur privacy problems. Revealing holistic information about private connections will usually lead to social dramas and embarrassment. The exposure of personal information will remarkably increase the chance of identity theft. The revelation of personal relationships will attract spamming and phishing. Social network sites may also misuse personal information for their commercial interests. Thus, users will always selectively reveal their personal information.

Given these benefits and concerns, users will attempt to maximise their gains from using social networks while protecting their privacies. There have been constant conflicts between publicity and privacy on SNSs. Common sense suggests that users will act on their best interest if problems emerge, usually in the cost of the global interest of the social network site. We will show in this chapter how current social network technologies

trigger the clash between publicity and privacy, which leads to friendship inflation, a serious problem that challenges today's social network.

3.2 Pride of Publicity

Social network publicity refers to the revelation of personal information including profiles and private relationships which are previously only fully available to the owners and their friends. SNS flattens the real-world social network by making personal information and social structure visible to users outside the ego-centric networks. The new social media is remarkably different from previous communication technologies such as emails and instant messengers, which simply provide one-to-one communications without reaching out beyond the ego-network circle. The publicity, though subject to some restrictions from users' privacy settings, can increase online social capital, improve the diffusion of knowledge and information and enhance trust and security on the Internet.

3.2.1 Social Networking

Social networks enable people to discover new friends and establish new connections through a chain of existing friends. It is assumed that people publicise their contacts of genuine friends, making them accessible by other users under some form of privacy restrictions. Users can browse profiles of friends of friends and trace the acquaintances' credibility based on the connections of mutual friends. In fact, Friendster, when launched in 2002, had leveraged a wide variety of contacts as meaningful connectors and recommenders for online dating[16]. This is significantly different from previous CMC methods because it flattens the social network and therefore the structure of social relationships are mostly visible to users. Unlike IRC (Internet Relay Chat) that facilitates anonymous chats, SNSs promote a sense of trust in users' social activities by encouraging the use of real identity. The genuine identities can be further guaranteed by other users who are directly connected to them. SNSs make vast amount of profiles searchable, so that users can locate other members with shared interests. People use SNSs for personal and professional use, communications, establishing new business developments and contacts. On SNSs, it is easier to join and connect to new people or communities within a similar geographic area, and to share common interests and join various urban tribes[101].

3.2.2 Dissemination of Knowledge and Information

There exist hubs in the social network which are more connected than other nodes. They are the centres of knowledge sharing and information inflow. Many of them are opinion leaders who can affect other people. The theory of assortativity suggests that

the opinion leaders, though in different knowledge domains and territories, will connect with each other closely. This forms a "reservoir" of information and knowledge that can circulate around the whole network in a short period. The process of information spread in the social network can be accelerated greatly if these people are specifically targeted[97][32]. In real world, these highly influential people may not be easy to be discovered and reached. On SNSs, however, as people publish their lists of friends, it is much more convenient to identify more active users by looking at how many contacts they have. Research has shown that SNSs can enhance information and knowledge sharing and viral marketing[84][26]. A notable example is that the South Korea-based social network site, Cyworld, which provides blogging, music and video sharing, was able to claim to have more traffics than the highly touted YouTube[1]. More singers and artists prefer to promote their music albums via MySpace to leverage the power of viral marketing. Many websites attempt to exploit the word-of-mouth strategy by adding social networking features.

3.2.3 Accountable Internet

People can easily publish information and share opinions on the Internet, yet it is not difficult for them to disguise themselves by remaining anonymous. As the Web keeps growing in a very fast pace, there are a huge amount of websites of different types that will produce tons of information and stories, some of which may well be rumours. It can be difficult to tell the rumours from facts in different expertise domains which the readers are not familiar with, not to mention the slightly exaggerated stories and deliberately bias views that appear more subtle and undiscernible. The online anonymity makes it difficult to hold people responsible for their activities and behaviours. This will cause many ramifications and problems such as spamming, malware, online security, trust and privacy concerns. Therefore, there have been proposals to argue for the establishing of social Web based on the existing Web and Internet infrastructure[72][106]. The idea is to bring trust and security to Internet by leveraging the peer-to-peer pressure on individual user[70]. The accountability can be achieved on SNSs as public pressure can be formed due to the difficulty to remain anonymous, for users generally publicise the connections to their real-world friends. Friendster, with its social network reach of four degrees when it was launched, is one of the first dating sites to take advantage of the publicity of profiles and contacts to provide trust and security.

3.2.4 Summary

Given the advantages that publicity can bring to the existing Web and Internet infrastructure on social capital, information sharing and publishing and trust and security,

---

[1]Cyworld News: http://www.usnews.com/usnews/biztech/articles/061109/9webstars.cyworld.htm

social network sites see a massive boom following Friendster's initial success in 2003. Friendster held the view that users should publicise their genuine profiles and contacts, otherwise the social graph will be devalued. The view is shared by many other SNSs such as Facebook, which encourages the use of real identity for online social networking. It is the belief and hope of system designers that the publicity of individuals' genuine information will benefit the individuals and the whole ecosystem of the social network.

3.3 Prejudice of Privacy

Privacy is the freedom from undesirable intrusions and the avoidance of publicity. Sharing personal information with not only friends but also acquaintances and even strangers will likely cause the leak of privacy. The common problems regarding privacy issues on using social network sites include exposure of backstage information, identity theft, spamming and phishing, and misuse of personal information. When SNSs were first introduced to users who are willing to adopt new technologies, they may pay much less attention to the privacy problems. But as more and more issues emerge from social networking, many will restrict their privacy settings to protect their privacy. In this section, we will discuss some of the major privacy concerns as follows:

3.3.1 Exposure of Backstage Information

When Goffman used the metaphor of theatre to explain people's social behaviours, he distinguished the *front stage* and *back stage* behaviour. Back stage behaviour is where performers are present but audience is not[51]. For social network sites, we use the term *backstage information* to denote the social information which users do not want to publicly articulate. For instance, they are not going to list their enemies and foes in public list. Neither will they express their dislike to some groups of friends publicly. It is rare for users to explicitly declare the ending of relationships which have already decayed. On the other hand, we will constantly adjust our behaviours according to different context. On Friendster, for example, users fear the presence of their employer and mother, as they mainly used the site for online dating. In MySpace there used to be a service called "Top 8" (now Top 40) where users can list as many as 8 close friends. However, when the list was full, users would either stop adding any new friends to the list, or replace the names in the list with those of new close friends without telling the old ones. The examples suggest that while exposure of backstage information will cause social dramas in real life, it will have similar effect in the virtual world.

3.3.2   Identity Theft

Unlike other functional methods for identity disclosure, users reveal more personal information on social network[115]. The revelation fosters a subjective and holistic disclosure of user's identity. The low entry barrier to social network sites and rich resources of personal information expose users to substantial risks of identity theft. Details such as contact address, age and date of birth are all potentially open to abuse. In networking sites such as Facebook, which users perceive as a more trustworthy place due to the presence of their real-world friends, more information about personal identity can be found and potentially misused. On the other hand, as people normally use social networks for keeping contact with their friends, they generally do not make the privacy settings particularly high. There might be many people aware of the privacy issues, but only a small fraction of them will change their default privacy preferences[57]. The less conservative privacy settings, coupled with the convenience of establishing "friend" connections, make SNSs more vulnerable for identity theft. It has been shown that it is possible to steal one's identity through widget applications on Facebook[2].

3.3.3   Spamming and Phishing

Most social networks require confirmation of friend request before one can access others' profiles and have closer interaction. However, due to the easiness of befriending, spammers can establish thousands of connections in a short period with only a handful of clicks. The theory of *Strength of Weak Ties* suggests that users are unwilling to aggressively reject most friend requests. It is not unusual for users with many friend requests to quit the site or heavily restrict their privacy settings. Some believe that SNSs have responsibility to stop spammings. Many sites attempt to curb spammings by resorting to technology and legislation. The technical methods include detecting scripts running on SNSs that attempt to harvest contacts. Recently, MySpace succeeded in winning the legal battle against one of the biggest spammers on the site[3]. Such methods are case-by-case and therefore can be inefficient. On the other hand, there is limitation on technology side. Email filters stop the junk mails by making the assumption that unsolicited messages about information like medication, penny stocks, fake university degree and software discounts are universally undesirable[128]. This assumption is no longer true in SNSs as they can facilitate unsolicited communication.

SNSs are an attempting place for phishers, who seek to acquire sensitive information from a victim by impersonating a trustworthy third party. Social networks recognise "circle of friends" which allows a phisher to harvest large amounts of reliable social network information[67]. Therefore, to mine information about relationships and common interests in a group or community, a phisher needs only to look at one of a growing

---

[2]http://www.bbc.co.uk/consumer/tv_and_radio/watchdog/reports/internet/internet_20071024.shtml
[3]http://mashable.com/2008/06/16/myspace-versus-scott-richter/

number of social network sites, such as Facebook and MySpace. Phishers can befriend with friends of targeted users only by online social grooming. By fostering mutual friend connections phishers can win the trust from targeted users and acquire their personal information.

3.3.4   Misuse of Personal Information

Users reveal a lot of information in the social networks by blogging, photo sharing, messaging, posting, etc. Some sites state in their privacy policy that they may provide personal information to a third party in order to facilitate or outsource some aspects of their services. Statements like this can be easily ignored by users as they usually do not read the terms and conditions carefully when registering with the site. Information sharing with third parties might provide better services to users, as always claimed by many sites, but it equally incurs risk of privacy leaking. Personal information can be misused by companies to facilitate their commercial interest. It is not uncommon that social network sites will exploit user profile information to mine data for targeting specific advertisements. Personal information related to consuming behaviours is particularly in great interest of advertising and marketing industry. For instance, companies such as Coke, Apple Computer and Proctor & Gamble are using social networking sites as promotional tools[12]. Another notable example is the failure of Facebook Beacon application. Facebook Beacon is an advertisement system that allow users to share their activities and behaviours, particular those about online purchasing, with their friends. The applications aims to leverage the power of viral marketing but due to its serious privacy concerns from the Facebook users, Facebook decided to drop the application after the protests from its users.

3.3.5   Summary

Given the issues discussed above, many users show their concerns on privacy when using social network sites. A recent survey of over 2,000 Americans, for example, found that 84% of Internet users worry about business and people they do not know getting personal information about them and their families[47]. While SNS designers hope that the users can publish as much information as possible, users will usually selectively reveal their information, acting on their best interest.

3.4   Public Exhibition of Private Connections

Social network designers require members to publicly articulate their private social connections. The public articulation is dramatically different from private description in

that they are supposed to be seen by users' friends and even users well beyond one's ego-centric network. Contemporary SNSs use a technique which may be called *static link* to represent the connections between the members. In this section, we will discuss how the technique triggers the clash between publicity and privacy, leading to friendship inflation.

3.4.1 Static Link

To befriend with someone, a user will typically have to send a friend request to him/her, as shown in Figure 3.1. He or she will need to confirm the request before the persistent connection is established, as shown in Figure 3.2. The connections will then be displayed on users' respective profiles. We called this befriending process the technique of *static link*. They can also choose to hide the connections by setting the privacy preferences. The technique guarantees the mutual recognition of the relationships between members in the social network. Some sites such as Friendster and Facebook provide functional descriptions of the relationships such as relatives, classmates and colleagues, in case that users want to elaborate the nature of the connections. Some sites may even impose an upper limit to the maximum number of friends a user can add. For example, the maximum number of friends used to be 1,024 on Orkut and 5,000 on Facebook[87]. Users can terminate old connections for new connections if the number of users' friends go beyond the limit or users confront with broken relationships.

FIGURE 3.1: Snapshot of Sending a Friend Request on Facebook
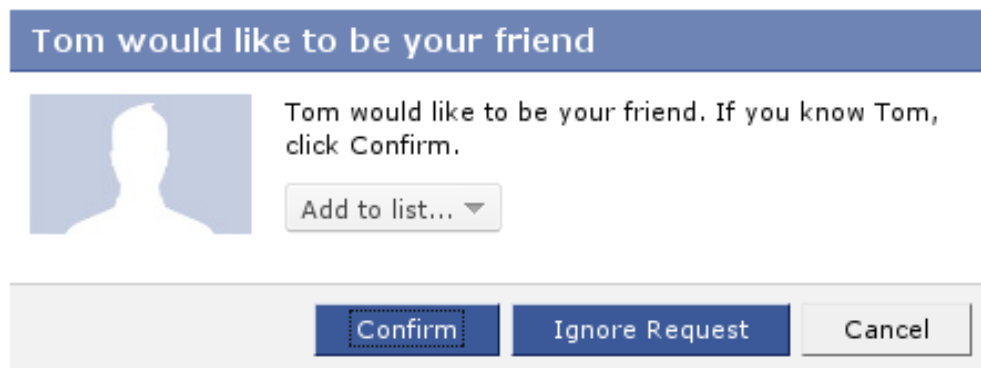
FIGURE 3.2: Snapshot of Receiving a Friend Request on Facebook

One can also categorise their friends based on the nature of relationships such as acquaintances, common friends, good friends, best friends and top friends on social network sites. The categorisation can be set private only to the owner. It can also be set visible to others. The categorisation may not be mutually recognised because no confirmation is needed from the users being described.

3.4.2 Friendship Inflation

Friendship Inflation refers to the online practice that users will usually acquire much more friends on SNSs than they can actually achieve in the real world. The phenomenon is first documented by Boyd on her study of Friendster. Friendster users' lists of Friends include fellow partygoers, people they knew (and people they thought they knew), old college mates that they had not talked to in years, people with entertaining profiles, and any one that they found interesting[17]. Danah Boyd noted that while some people are willing to indicate anyone as friends, and others stick to a conservative definition, most users tend to list anyone who they know and do not actively dislike[16]. As a result, some people use the term "friendster" to signify the acquaintance of casual connections[16]. The problem is echoed by Fono et al. in their research on LiveJournal. They coined the term *Hyperfriendship* to indicate the differing and multiple views of what "friendship" means. For Facebook, a 2006 research reported that the average number of friends was 272. A 2008 survey suggested the number rose to 395. Noted that Facebook could

only be registered with a dedicated university email before it opened the registration to public in September 2006.

There are several causes and motivations for users to inflate their friend lists. Research suggests that friendship inflation can be partially explained by the multiple meaning and different interpretation of friendship. On Livejournal, Fono and Raynes-Goldie found that there are seven interpretations of friendship[46]:

- Friendship as Content: being a friend can access others' journals.
- Friendship as Offline Facilitator: choosing friends according to whom they know offline.
- Friendship as Online Community: leveraging the strength of weak ties by forming online communities.
- Friendship as Trust.
- Friendship as Courtesy.
- Friendship as Declaration:
- Friendship as Nothing: friending someone is merely a matter of adding another name to a list.

Boyd examined friendships on Friendster and MySpace and found the following reasons for befriending with one another on social network sites[17]:

- Actual friends
- Acquaintance, family members, colleagues
- It would be socially inappropriate to say no because you know them
- Having lots of Friends makes you look popular
- It's a way of indicating that you are a fan (of that person, band, product, etc.)
- Your list of Friends reveals who you are
- Their Profile is cool so being Friends makes you look cool
- Collecting Friends lets you see more people (Friendster)
- It's the only way to see a private Profile (MySpace)
- Being Friends lets you see someone's bulletins and their Friends-only blog posts (MySpace)
- You want them to see your bulletins, private Profile, private blog (MySpace)
- You can use your Friends list to find someone later
- It's easier to say yes than no

As shown above, social network users befriend with others for many reasons and purposes, some of which are simply due to the convenience of the technology. However, even we assume that all members befriend with those who are genuine friends in the offline world, just as what most SNSs encourage their users to do, the number of publicly listed friends will still keep growing. In the beginning, suppose there are n friends in a user's public list, when she acquires the $(n+1)$st friend in the real world, she will add him/her to the list. This friend may be one of her classmates, colleagues and neighbours. When n becomes very large with a significant part of the contacts having not been used for a long time, she may want to "clean up" the list. But because the list is publicly displayed, to remove anyone from the list will expose her backstage information. She will risk to offend the people being removed by publicly declaring that they are no longer in her

friend list. The public declaration can also be seen by other users, which causes further social implications and ramifications[91]. Fear of rejection and removal pushes up the number of n. In fact, the demand from users to increase friend list size has forced Orkut to lift its 1,024 limit and Facebook to remove the cap of 5,000 friends limit. Figure 3.3 shows a warning from Orkut when users have more friends than the upper limit it has set.

FIGURE 3.3: Orkut's Friend Limit

Some sites attempt to mitigate the problems by providing objective descriptions for connections such as relatives, classmates and colleagues, only to find that members routinely ignore the descriptions. The fact is that many users do not bother to add more details of their friendships. Wherever users do utilise the categorisations to describe their friends, it does not effectively mitigate the problems. While the number of relatives and classmates may remain stable, the number of other categories such as friends and colleagues will generally keep increasing. Further, the descriptions may not reflect the closeness of the relationships. For instance, "Went to school together" relationship may well be perceived to be closer than "Through a friend". However, this may be incorrect if friends in the former case do not contact with each other after graduation but friends in the latter case keep close contact on a regular base. Worse still, as users' network evolves over time, the connections will constantly change. But users are not keen to update the connections in the system until necessary.

3.4.3   Top Friendship Inflation

To curb the friendship inflation, some social networks introduce top friends application. As the name suggests, the application provides a tool for users to select some of their best friends from the bloated friend list. It should be noted that the top friendship needs not be mutually recognised. That is, if A adds B as her top friend, she does not necessarily inform B about the move. In fact, in most case, A will just keep B uninformed. The privacy setting of top friends can be set to be private. Alternatively, it can be set to be visible to other users. Based on our definition of social network, we may only focus on the relationships that are mutually recognised and therefore we investigate the case when it is set to be public.
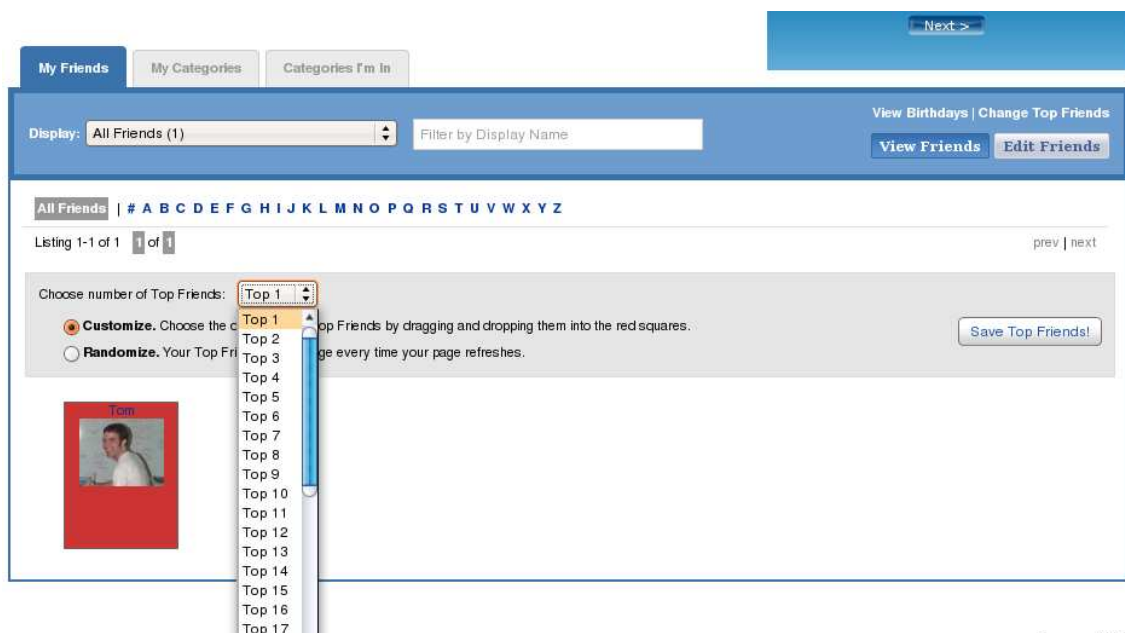
FIGURE 3.4: Snapshot of MySpace Top Friends Management Interface

On MySpace, this type of service was originally called "Top 8", allowing only 8 best friends to be listed. With demand from users, the number gradually enlarged to 16, 24, 32 and now 40, as indicated in Figure 3.4. On Facebook, the relevant application is called "Top Friends". It imposes an upper limit of 32, which is shown in Figure 3.5. However, users constantly request the increase of the friend listing space. In fact, the call for increasing space is one of the most discussed topics on the official discussion board of "Top Friends" application. Given the need of a bigger friend list, there is another Facebook application called "Super Top Friends" (later renamed as "My Top Friends") offering a maximum number of 64, double the number in "Top Friends" application.

Top Friends application requires users to demarcate the border between top friends and non-top friends. While it is easy to add someone to the list, removing someone is another story. Because the list is publicly accessible, removing people from the list usually leads

to many social implications. The social damage can even be greater than removal in the ordinary friend list. This is because the name of "Top Friends" implies a much closer connection than the name of "Friends" does. Degrading from "Top Friends" to "Friends" is therefore conveying a more devastating message to that from "Friends" to "Non-Friends". Whenever a user attempts to shrink her "Top Friends" list, she may seek to avoid informing the people to be removed by changing the privacy settings. This is particularly true on sites like Facebook, which provide a fine granular control of privacy settings than some other sites. However, the change of visibility also changes the nature of reciprocity of the relationships. The "Top Friends" connections are no longer mutually acquired and recognised. It will end up as the result of one's favourite friends in her private address book. On the other hand, users who wish to keep the list public often force the websites to push the friend upper limit higher and higher. Given the lesson from Friendster, which did not listen much to what users said, many SNSs such as MySpace and Facebook would yield to the pressure from users and increase the upper limit of "Top Friends" list endlessly. Thus, "Top Friends" application repeats the same consequence of *static link*, leading to top friendship inflation.
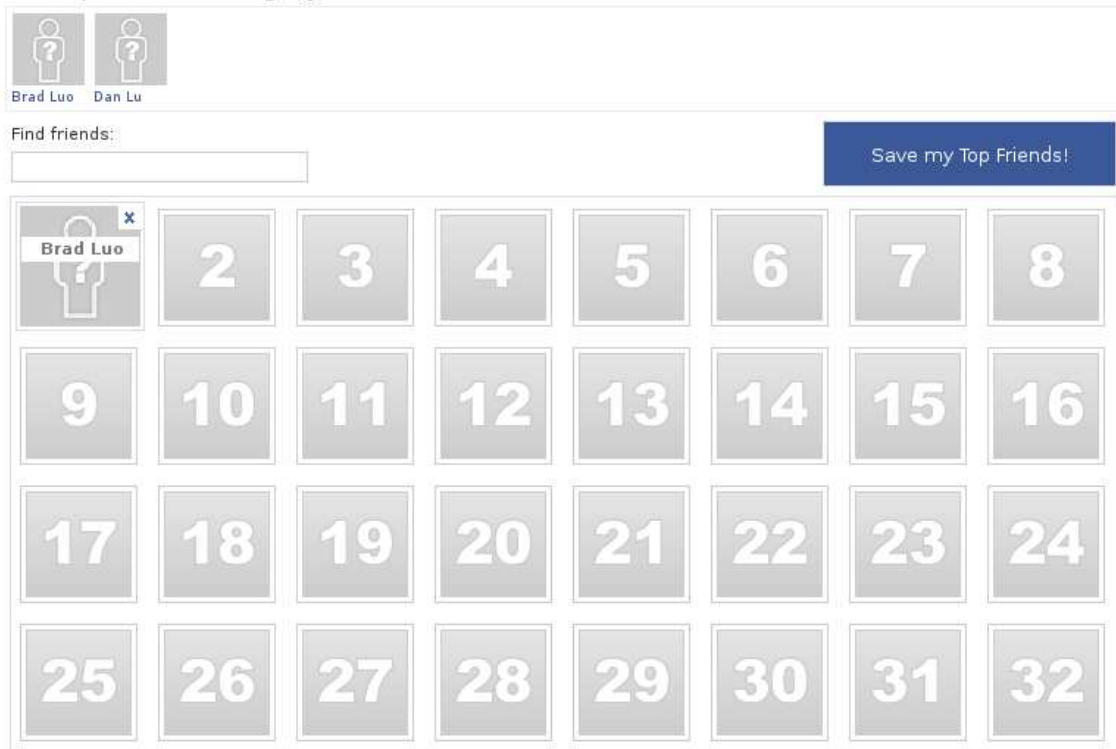
FIGURE 3.5: Snapshot of Top Friends Application on Facebook

3.4.4 Friendship Collectors

For the two cases mentioned above, the number of "friends" keep increasing but in a relatively slow and stable pace. Most members are just ordinary users who use social networks for maintaining the established social network and expanding it gradually based on their offline activities. However, because the connection is based on self-description with near-zero cost of befriending, some members will exploit the convenient technology by collecting as many contacts as possible, regardless of the number of genuine friends they actually have in the real world. A notable example is one of the Facebook applications, called PackRat[4], where users collect sets of illustrated cards for points and levels[5]. One part of the game is to "Steal from Friends" in order to find the right card to complete a collection. The game encourage users to befriend with as many other users as possible so that their cards can be obtained. The practice of "befriending just for gaming" is strongly criticised by Facebook, which, like Friendster and other social network sites, is eager to promote real-life identity and genuine friendships. But because it is perfectly legitimate to befriend with anyone on the site, many social networks do not offer any technical means to prevent users from befriending with a large amount of other users in a short period of time.

Boyd identified the first generation of friendship collectors on *Friendster*. They were called "Friendster whore", who aggressively stocked up superficial friends. There are several reasons for the emerge of "Friendster whore". First, *Friendster* only allows users to access those within four degrees[16]. If a user wants to browse more profiles, they have to add more friends. Many users who attempt to enlarge their dating portfolio will collect as many contacts as possible. Second, there is a popularity rating on *Friendster* which is higher if the number of friends is bigger. The higher the rating one has, the easier he or she can be searched for by other users. Third, people might befriend with someone simply for political reasons. If they see the value of putting their friends in certain category, they may try to manipulate the list by deliberately collecting contacts. The large amount of friend connections in turn makes the friendship collectors appear popular, which forms a positive feedback. Finally, spammers and phishers will take advantage of SNSs to spread or collect information about users' behaviours and activities[128][67]. These are usually technically advanced users who can rip off the data from social network sites by running programming scripts. Some scripts are so powerful in taking up the servers' computing resources that many SNSs will regulate the use of the social network by monitoring the traffic of the site.

Contacts may also be passively acquired. There are people who themselves are highly popular offline so when they publish their profiles on SNSs, they receive a lot of friend requests. The nickname "Facebook whales" vividly describes the prestigious group of

---

[4]http://www.alamofire.com
[5]http://www.techcrunch.com/2008/09/15/facebook-isnt-a-social-network-and-dont-try-to-make-new-friends-there

Facebook users who are able to accrue much more connections than average users[69]. In fact, these people are usually bloggers, journalists and celebrities. The case is different from the previous one in that they act as the hubs and opinion leaders of the site. They usually make a positive contribution to SNSs by attracting their fans to use the sites. However, it is exactly because both of these beneficial friendship collectors and malicious ones share similar behaviours in terms of having large amount of contacts that social network sites are not able to distinguish them.

The existence of friendship collectors indicates the vulnerability of most social networks. It signals the fundamental weakness of *static link*. As long as the friendship establishment is purely controlled by users at near-zero cost, connection is always subject to abuse by deliberate collectors. There is no guarantee of the quality of friendship in an ever-growing public social network sites.

3.4.5 Fakesters and Fraudsters

The term *fakesters* originated from the early social network site, Friendster. They quickly become popular in the following SNSs. Fakesters are fake personas created by users for different purposes. Figure 3.6 shows a fakester, Tony Blair and his fakester friends on MySpace. The profiles of these political leaders and celebrities are purely constructed by ordinary users, yet these fakesters often connect with other fakesters and entwine with the rest of social network. By connecting to real people they become an integral part of the social network.

There are different types of fakesters, which reflect users' social and cultural characteristics. Research on Friendster, for instance, revealed three categories of fakesters[16], as shown in the following table. Here, both cultural characters and community characters can be served as a unified symbol under which real users can connect to each other with similar preferences.

1. Cultural characters that represent shared reference points with which people might connect (e.g. God, George W Bush);
2. Community characters that represent external collections of people to help congregate known groups (e.g. Brown University, Black Lesbians);
3. Passing characters meant to be perceived as real.

The presence of fakesters is increasingly a commonplace for most SNSs, particularly those that can be registered with a public email address. Anecdotal analysis of profiles on Friendster, MySpace and Orkut shows the ratio between the authentic profiles and the fake ones[89]. Table 3.1 suggests that the ratio between authentic profiles and fakester profiles is 20 to 10 on Friendster, 23 to 7 on MySpace and 29 to 1 on Orkut. Note that
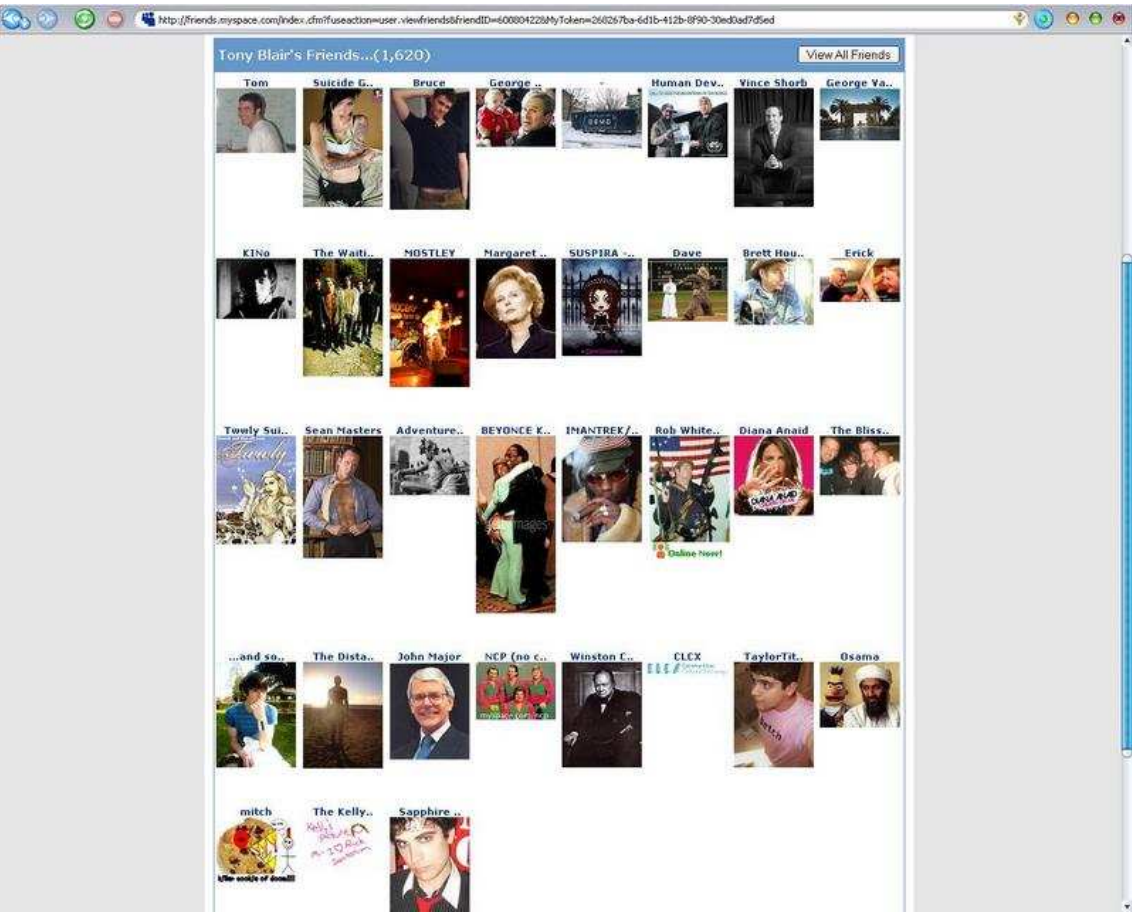
FIGURE 3.6: Fakesters on Myspace: Tony Blair's Friends

Orkut could only be registered by invitation whilst both Friendster and MySpace could be registered using a public email address when the data was collected.

Facebook opened to public registration since September 2006[57]. Within two years, fakester profiles have increased substantially. This can be seen from the ratio between real names and fake names, as shown in Table 3.2. It shows that the percentage point of Facebook fake names is 8%, compared against 89% of real names and 3% of partial names.

The fakester phenomenon reflects the dynamics of the users. Social network users are extremely active in creating fakesters. An important and practical motivation for them to create fakesters is to broaden their network reach and look for like-minded people[17].

TABLE 3.1: Authentic vs. Fakester Profiles

SNS	Authentic	Fakester
Friendster	20	10
MySpace	23	7
Orkut	29	1

For example, people connecting to the same fakesters of cultural characters and community characters may share similar social and cultural preferences. By the same token, people who admire the same celebrities may have common language in certain area. Because of the popularity of fakesters, some attractive fakesters can make a lot of "friends", which boosts friendship inflation greatly. The highly popular fakesters may be deliberately constructed by spammers who want to harvest contacts for their special purpose, but they can also be created by ordinary users who simply want to have fun. Fakesters can generate a great number of fake connections, making a substantial part of the whole social network. Therefore, many social network sites strongly discourage the creation of fakesters. They argue that these fake profiles will collapse the structure of the network and devalue the meaning of connections between people. Unfortunately, they do not have any effective technology to distinguish fakesters from real users, casual connections from close connections. Some companies, such as Friendster, had attempted to eliminate all of these fake users by directly removing them from their sites and servers. This affects the creativities and activities of many users. Most users who create fakesters are not spammers and do not seek to devalue the social network in the first place. The undiscriminating removal, however, created tension between the company and users[16]. As a result, the company saw massive rebellions from the users.

3.4.6 Summary

On SNSs where links are public exhibitions of private connections, the *static link*, which assumes that the cost of social grooming is near zero, triggers the clash between publicity and privacy, leading to friendship inflation. The phenomenon of friendship inflation and top friendship inflation implies that the problem will persist as long as the site continues to use the static description method. The lack of effective technology to cope with fakesters and fraudsters complicates the issues. While both the site and users will not get benefits from friendship inflation, it seems the site suffers most.

3.5 Public Display of Private Self

Social network sites typically provide profile services for users to present themselves. System designers hope that these profiles can form an array of individual identities that are consistent so that they can be discovered and searched more efficiently by other

TABLE 3.2: Fake Profiles on Facebook

Category	Percentage Facebook Profiles
Real Name	89%
Partial Name	3%
Fake Name	8%

members. However, the profile services ignore the need for users to present to different audience with different information. With more relatives, colleagues, neighbours and many other different types of friends participating in the social network, a universal profile simply fails to adopt to different context. We will discuss in this section how universal profile produces generic persona, which causes social embarrassment as social network grows.

3.5.1 Universal Profile

Most social network sites allow people to present themselves through profiles. The profiles were very simple in the first generation of social network sites such as *SixDegrees.com* but since Friendster, they have become increasingly rich in description, thanks to the advancement of Web technologies and standards. A profile usually includes but is not restricted to name, birthday, location, hometown, interests, education and work history. Some may also display information about their social networks, relationship status, contact methods, etc. Some of these options are enumerative, such as gender and political views. This means users can only select one from the pre-defined list. Many options can be filled in with limited number of characters, some of which can even utilise the features of HTML and Javascript language. It is commonplace for users to share photos and videos on their profiles. Figure 3.7 shows a profile from Facebook. Profiles usually employ WYSIWYG (What You See Is What You Get) method so that the editing effect can be seen immediately. Users can change the permission on which part of the profiles are visible to whom. However, profiles are viewed with no difference by all visitors. It can not detect the visitors based on the nature of the connections such as parents and employers. Thus, the profile is generic and universal on the social network sites.

Profiles are essentially public display of private self. Because of the privacy of backstage information, users always selectively reveal their information to their friends, even the close ones. They will keep their behaviours coherent and adjust to different context[51]. However, the universal profiles force them to present to different friends and people with the same content.

3.5.2 Generic Persona

The profiles represent what the users choose to present their identities. As the users and their networks grow over time, their profiles may change correspondingly. The profiles reflect users' online personas. Users usually put the best effort to make the profiles represent themselves as accurately as possible. For example, a research suggested that users on Facebook "reported high confidence that their Facebook portrayals described them accurately and that those portrayals were positive"[79]. But it is not unusual for people to take photos of celebrities and put them on their profiles.
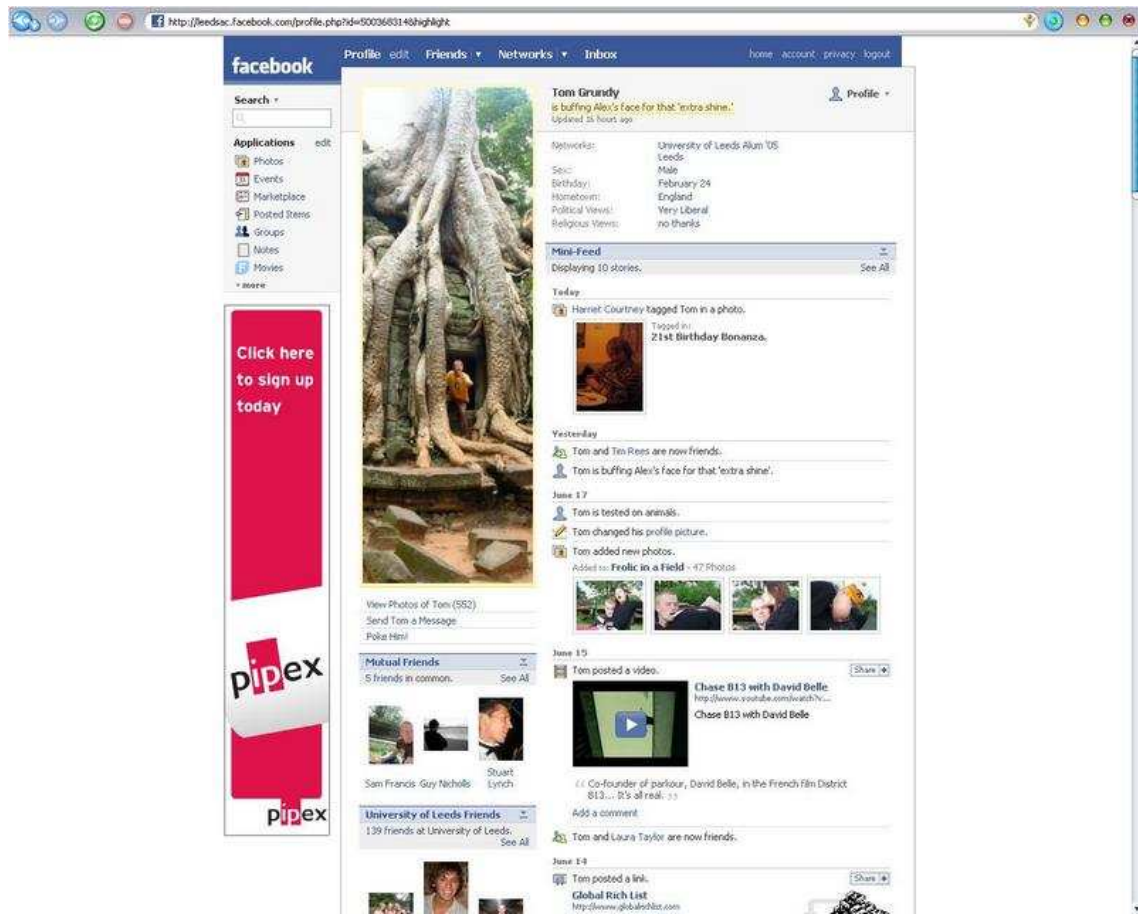
FIGURE 3.7: The Universal Facebook Profile

In our daily life, we usually present ourselves differently to different audiences and we attempt to behave appropriately in different situation and context[51]. When social network sites move people's relationships to cyberspace, users bring their various social masks online. However, most sites do not yet provide tools for managing multiple profiles and masks. The communication goes well only when a specific group of people using the sites but it will cause problems when more users and audience from different real-world social groups participate. On Facebook, for instance, users view their audience as peer group members, but not faculty, administrators within the campus, or outsiders. Thus, they behave in a way similar to what they do in the student communities. This might be significantly different from what they do when talking to the faculty members. Therefore, some Facebook users feel uncomfortable when their profiles are viewed by faculty members. Facebook users do not have any choice but only one face on Facebook.

It has been shown on Friendster that most users fear the presence of two people: boss and mother[107]. Interestingly, teachers also fear the presence of their students. Social network sites address this issue by giving users control of their profiles by adjusting the privacy settings. As a result, close friends can see all of the profiles and others might just see part of them. This function may solve the privacy problem but do little to

the online persona. In real life, teachers, relatives and working colleagues are all close contacts of us. They know us very well. We are happy to communicate with them appropriately in different situation. Social network sites, on the contrary, are much less context sensitive. They are usually motivated by commercial interest and are eager to attract more users but fail to provide adequate tools to accommodate multiple online personas. The sites will become an increasingly embarrassing place for socialising when connections and interactions become more diversified.

3.5.3 Summary

Profiles are public display of private self. While we constantly adjust our behaviours to different setting and context, social networks do not provide us the same versatile tools to present ourselves. Users often end up with making their online personas more generic to avoid social embarrassment. Universal profile triggers clash between publicity and privacy, which leads to generic personas. As more people join the network and the relationships become diversified, users will be more cautious to present themselves because they have to take all their friends into account. While both the site and users will not get benefits from the lack of multiple profiles on the site, it seems users suffer most.

3.6 Discussion

Publicity and privacy are two fundamental forces that drive the development of online social network. Without publicity, users can not browse others' social network. Without privacy, users risk to expose themselves to strangers and spammers. A balance should be carefully negotiated between system designers and users. Unfortunately, the technologies currently employed by most SNSs such as *static link* and universal profile brutally damage the delicate balance. They trigger the clash between publicity and privacy, leading to friendship inflation and generic personas. While the problem of universal profile may be remedied by providing multiple profiles and communication channels, there are no easy solutions to friendship inflation, as we show before. However, friendship inflation causes far more damage to the integrity and usefulness of the social network. We will discuss the consequence of friendship inflation in more details in the next chapter.

Chapter 4

Hyperfriendship Social Network

4.1 Introduction

Inspired by Baudrillard's notion of hyperreality[14], the term *hyperfriendship* is first used by Fono et al.[46] to describe the differing and multiple views of "friendship" on social network sites. We use the term *hyperfriendship social network* to describe the online social network with inflated number of friendship connections. We will first present statistical evidence on the macroscopic scale to support the microscopic analysis of friendship inflation. These include our data on more than 20,000 Facebook users in the University of Southampton. Then we analyse how the model affects information spreading and plays a negative impact on social networks. Finally, we review the rise and fall of some major social network sites.

4.2 Cumulative Network Model

Hyperfriendship social network can be understood as a cumulative network where edges are added and rewired without removal. Being a superset of real-world social network, cumulative network exhibits some interesting features such as no definite cutoff and dissortative mixing which distinguish itself from the topological characteristics of real-world social network. These features have been repeatedly found on many established social network sites. Our three-year observation on the evolution of the network of the University of Southampton on Facebook confirms the deformation of network topology over time. The topology of cumulative network has major impact on information sharing and dissemination.

4.2.1 Rewiring Without Removal

There are several factors contributing to friendship inflation such as friendship collectors, fakesters and fraudsters, spammers and phishers, as we analysed in the previous chapter. To be fair to most SNSs, we assume that SNSs are well policed and most members only add friends with whom they have actually met offline. Gradually, their offline social activities will bring more friends to their online networks. People have limited time and energy to maintain stable social relationships. In fact, there is a supposed cognitive limit to the individuals with whom people can maintain stable social relationships[37]. As a result, some of the old connections will gradually decay when we acquire new ones. On complex network study, this may be modelled as *edge rewiring*. Online social networks are capable to preserve old connections, leading to *rewiring without removal*, a character not existing in real-world social network but commonplace on social network sites. The effect of rewiring without removal of decaying connections is illustrated in Figure 4.1. Dashed lines represent decaying real-world connections that have been maintained as online social connections. Every time people make new contacts and leave some old contacts obsolete, the old contacts can always be preserved in the social network. As an SNS grows, its social graph will become denser and denser.

On SNSs, people are highly unlikely to explicitly declare the ending of any connections that gradually decay. The technique of *static link* employed by most SNSs requires users to articulate their friends publicly by demarcating the borders between friends and non-friends. Therefore, users prefer not to remove any fading connections to avoid offending people. Users also worry that the removal of unused connections will have many implications and ramifications that may not be predicted at the time of action. On the other hand, the popularity of Top Friend applications and services on social networks like Facebook and MySpace suggests that SNS owners seek to mitigate the problem of friendship inflation by "upgrading the relationships" rather than "downgrading them".

When many users have more connections than they actually do, the topology of the network will increasingly diverge from that of real-world social network. We propose a model to simulate the growth and evolution of the cumulative network.

The model is based on BA network as discussed in the second chapter. It has been observed that both conditions in the original model, *growth* and *preferential attachment*, apply to social network sites. In addition, two modifications and one condition are added to the model:

(a) In BA model, the exponent $\alpha=3$, but in real network, the number is between 2 and 3. We use 2.3, which is the measure for film actor collaboration network based on Internet Movie Database (IMDb).
(b) BA model does not specify the value of m, the average degree of the network. Dunbar's number suggests that people are capable to maintain regular contact with about
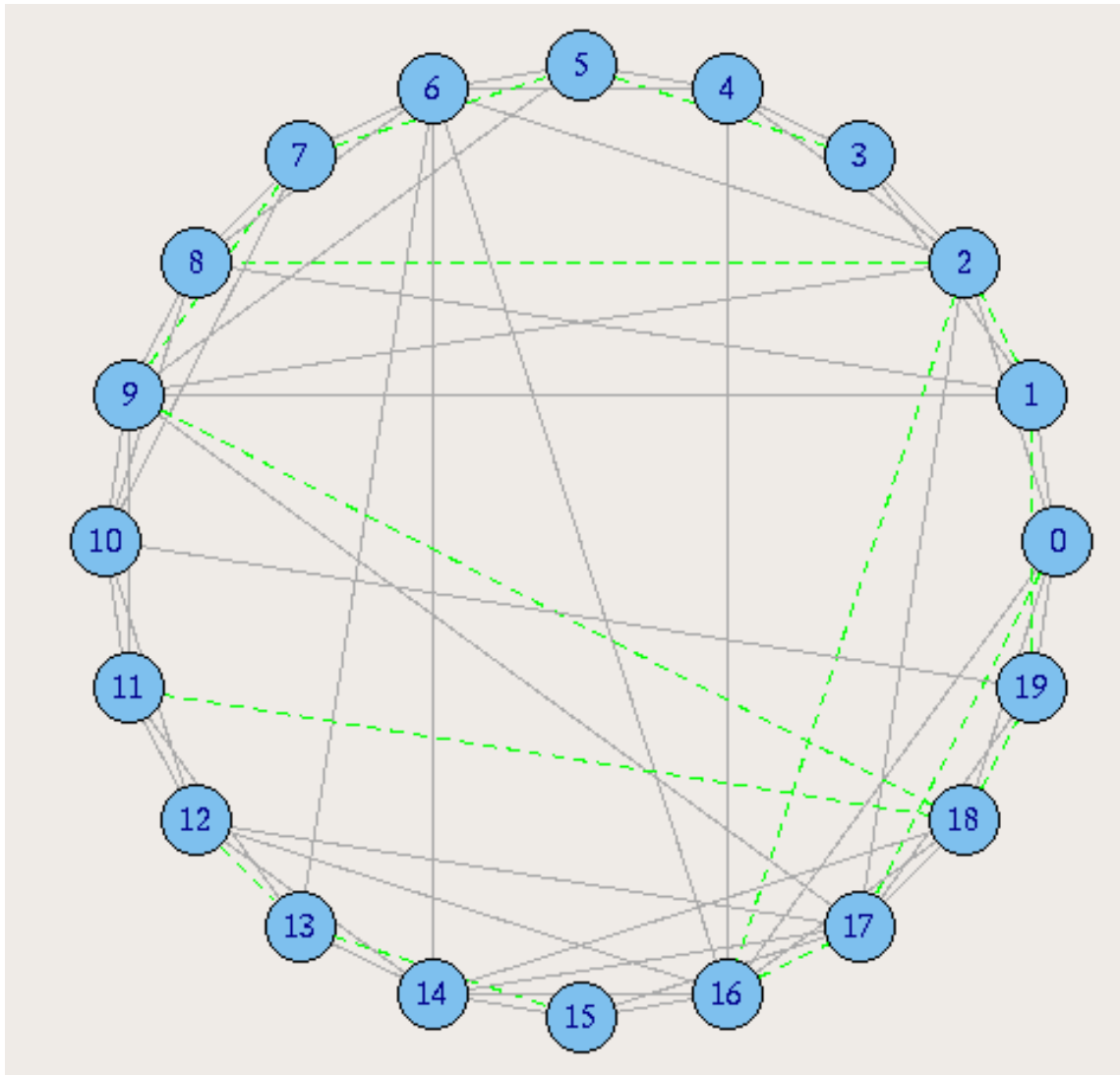
FIGURE 4.1: Illustration of the effect of rewiring without removal of decaying connections. Dashed lines represent decaying real-world connections that have been preserved as online social connections.

150 friends. The number can be interpreted as the lower bound number of links one can have. Therefore the value of m, which is the number of friends that people claim to have, should be no less than Dunbar's number. For our convenience, m is set to be 150. (c) Individuals will make new acquaintances and forget old links after joining the network. This is called *edge rewiring*. BA model does not take into account the effect of internal edge rewiring. We assume in our model that every node will rewire its m edges to other nodes with probability p_r proportional to d^{-r}, where d is the social distance (described in chapter 2) between them and r is an adjustable constant. This condition will only be used qualitatively in our model.

With only (a) and (b), we have a new function for probability p_k:

$$p_k = 2m(m+1)k^{-2.3} = 45300k^{-2.3} \tag{4.1}$$

Figure 4.2 shows the graph of Eq 4.1.
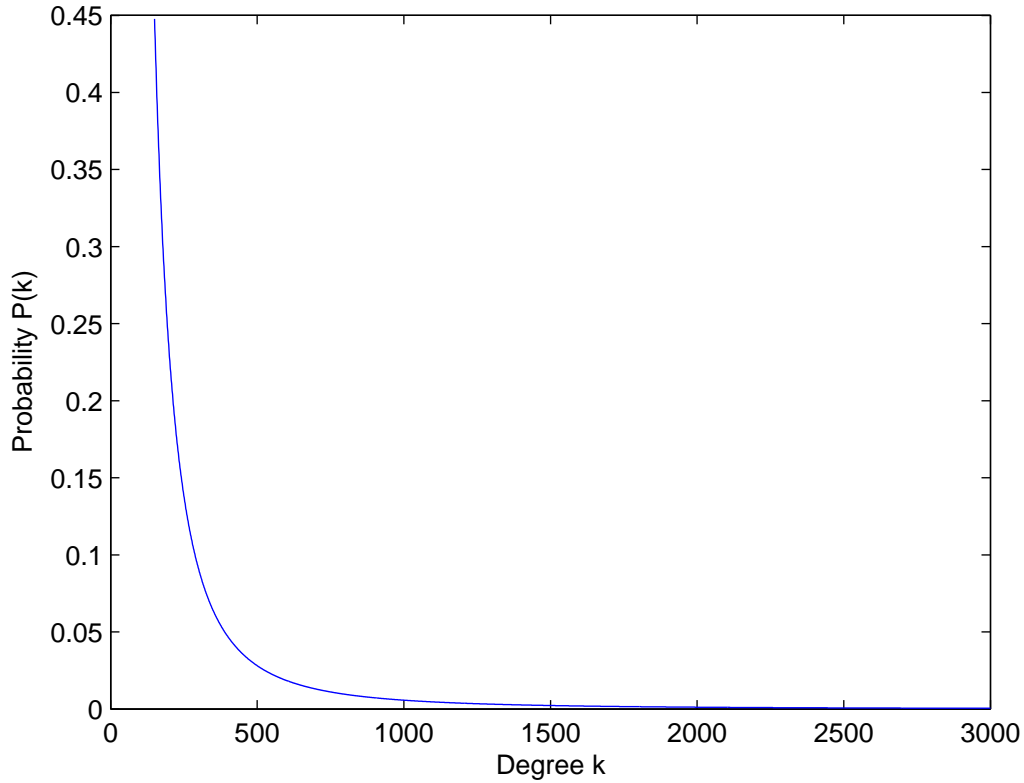
FIGURE 4.2: Degree Distribution in BA model with m=150, α=-2.3

The graph suggests that in a social network with m=150, about 44.78% of the people have about 150 friends. The remaining part of the population are able to maintain stable contact with more than 150 people. This is true regardless of the size of the network as it is scale-free. Notice here the notion of friends at least includes family members, neighbours whom you know and people whom you have worked or studied with for some time. As of this writing, empirical data shows none of the social network sites gain the percentage of 44.78% or above, indicating that people have not yet fully moved their real-world relationships online. However, as the social network sites grow rapidly in the recent years, we would expect the percentage will approach to that of the real-world network in a short period. Condition (c) suggests that people will "rewire" the friend links if they could not afford to keep regular contact with them, thus leaving a long trail of socialising footprints. In cumulative networks, the obsolete connections will not disappear automatically, which is contrast to the real social networks where

old relationships will decay gradually when people do not maintain a certain degree of social grooming with each other. We discuss two scenarios of the consequences for the development of social network sites:

Scenario 1: as the number of friends goes beyond 150 and continues to grow, it is not uncommon to find people who have hundreds of thousands of friends. In the real world, nevertheless, people with many contacts are usually the rich, the politicians, the celebrities and the leaders. Ordinary people may like to make friends with these high-profile figures, but usually find it very difficult to do so. However, on social network sites, the notion of high degree simply does not imply the high social status of the individual, as in the case of offline world. This will destroy the factor of *preferential attachment* as described in BA model: people now do not make friends by looking at their number of contacts. Model A of BA network shows that without *preferential attachment*, the network will lose scale-free character.

Scenario 2: if at some point, the network stops to grow, then the size of the network will remain unchanged or even shrink. This is quite common as social network sites stop growing and start losing the members for lack of attractiveness. Then members of the network can only make friends with other existing members. This simply increases the clustering coefficient of the network, making it a denser place. In the end, it will become a random graph with extremely high probability for an edge to be placed between any nodes. In particular, if people still keep making friends in the pattern of *preferential attachment*, the graph will exhibit a Gaussian distribution. In another word, the number of new friends are proportional to the number of friends already acquired, and this will keep doubling. In both cases, the network will lose the power law distribution of a scale-free network.

4.2.2 No Definite Cutoff

For real-world social network, nodes have a finite life time and finite edge capacity[5]. In the film actors' network, for example, elderly actors have less attraction to the young who newly join the network. The ageing factor is particularly important when discussing social network. It will affect the topology of the network such as power-law degree distribution, clustering coefficients and small average path. To address the issue, Dorogovtsev and Mendes proposed a network growth model which incorporates the effect of gradual ageing[35]. They proved that a reference network with ageing results in cutoffs of the power-law scaling, which fails to maintain the scale-free characteristics of complex network. Thus, as the network grows, it will gradually change the topology, showing a finite cutoff[35]. This may also be explained by the fact that individual has limited amount of time and energy to befriend with others. Thus, the scale-free character can not go on forever. Both Figure 4.3 and Figure 4.4 indicate when the ageing factor or capacity
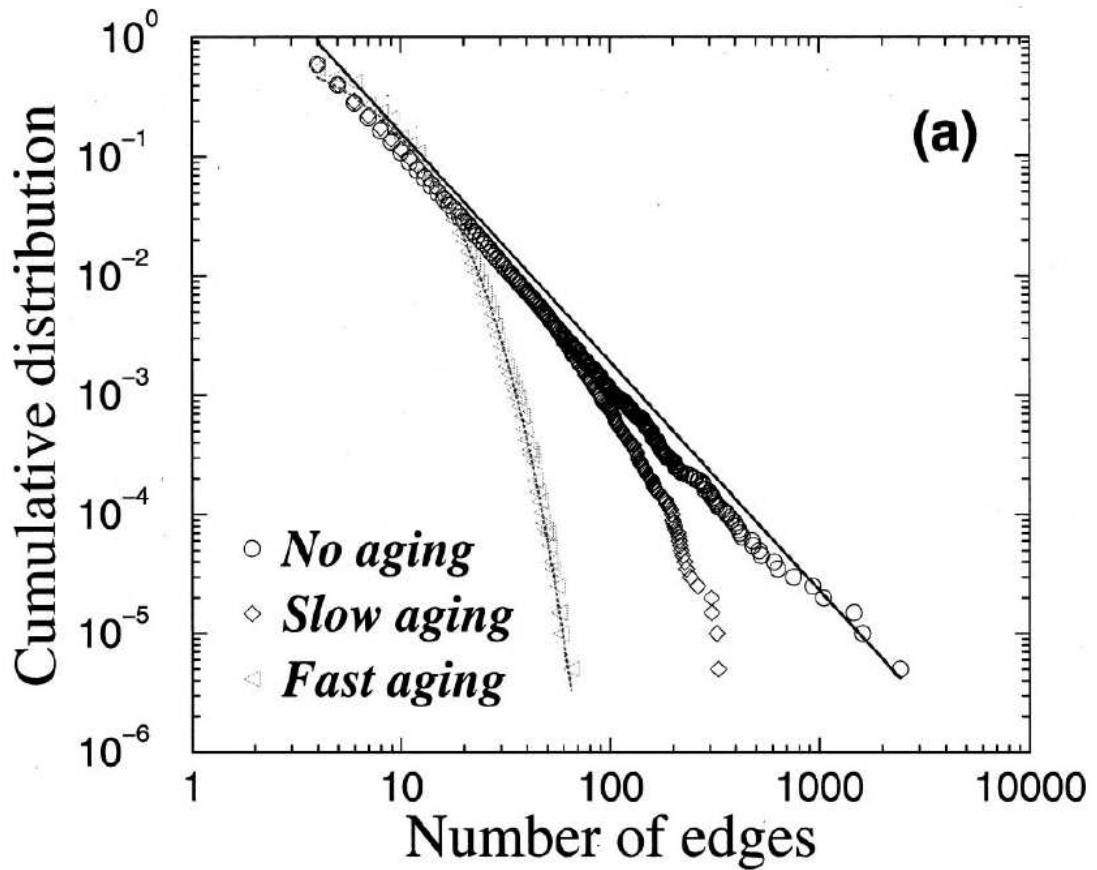
FIGURE 4.3: Deviation from a power law degree distribution due to adding age to the Barabasi-Albert model. The constraints result in cutoffs of the power-law scaling. Taken from [5]

constraint is added to the Barabasi-Albert model, it will result in definite cutoffs of the power-law scaling.

However, on a social network site, there generally lacks of a clear cutoff beyond the scaling region. We theorise that this is due to the effect of *rewiring without removal*. In the beginning, users register with the social network site and invite their friends who they think might be interested to join the site. As they explore the social network, they will acquire some friends. These newly acquired contacts can be people who share mutual friends with the users. They can be people who share similar interests and social and cultural backgrounds. They can also be people with whom the users come across with during their online social activities. These contacts form users' ego-centric networks. Because of the convenience of moving the offline connections online and befriend with new friends, these ego-centric networks will quickly become saturated, – a situation where users add a few hundreds of friends and reach their capacity constraints. However, the friend-making process in social networks is so cheap that users can continue to acquire "friend" connections with many more people if they want. The *static link*
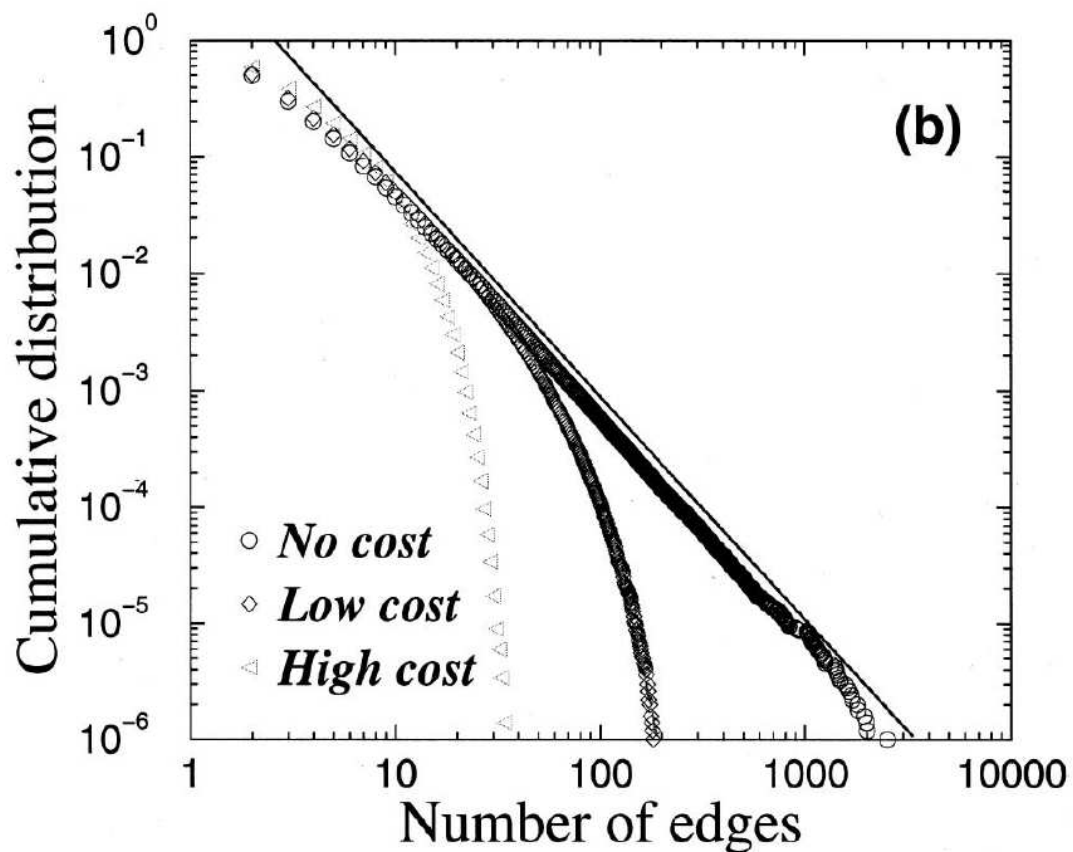
FIGURE 4.4: Deviation from a power law degree distribution due to adding capacity constraints to the Barabasi-Albert model. The constraints result in cutoffs of the power-law scaling. Taken from [5]

means these connections, once made, are permanent. There will not be any significant ageing effect in online social network. It is not unusual to see people with thousands of friends can add more friends on a daily basis.

The absence of a definite cutoff of the degree distribution of real-world social network has been repeatedly found by many research and studies on social network sites. Holme et al. investigated the structure and time evolution of an Internet dating community, *pussokram.com.* They analysed the contacts, friendship confirmations and messages, guest book and flirts on the site. They found that while the degree distribution is highly skewed, it is "interesting to note that there are no clear signs of the (inevitable) high-degree truncation in any of the graphs"[63], as shown in Figure 4.5. For Libennowell's research conducted on Livejournal, although the outdegree distribution exhibits a finite cutoff, the same patten does not hold true for the in-degree distribution[88], as indicated in Figure 4.6. The conclusion slightly contrasts to Mislove's research on the same site. They see no clear signs of cutoffs on both indegree and outdegree distribution[93], as shown in Figure 4.7. It should be noted that Mislove's data, collected in December 2006, covers 5.2 million users and 72 million links while Libennowell's data, collected in
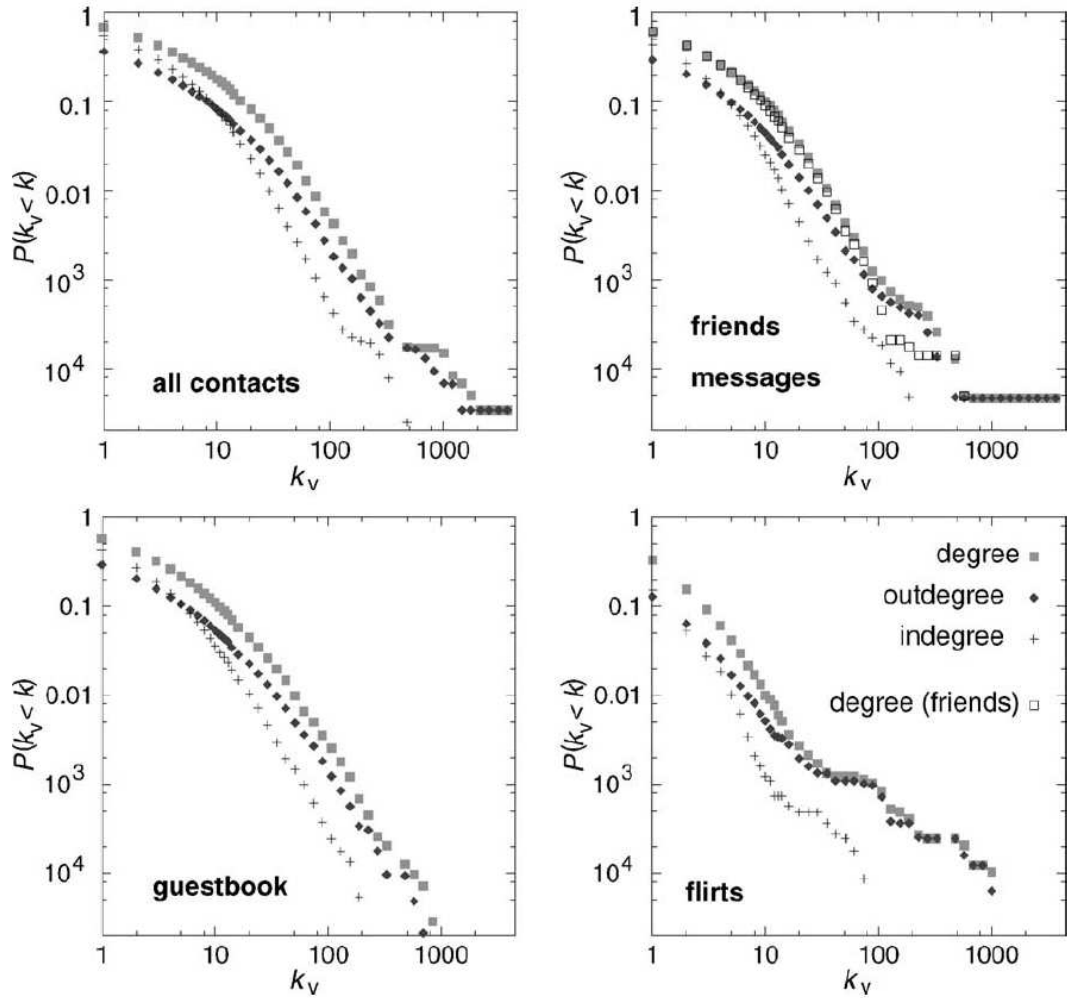
FIGURE 4.5: Degree distribution for pussokram.com. Taken from [63]

February 2004, covers 1.3 million users and 4 million links. The cutoff is also absent in the Japanese SNS, Mixi[83], as indicated in Figure 4.8. Ahn, Han and Kwak investigated the degree distribution for both Cyworld and MySpace[4], as shown in Figure 4.9 and Figure 4.10, respectively. None of them demonstrate a clear cutoff.

4.2.3 Dissortative Mixing

Users are able to befriend with different people from across different groups on social network sites. Even after they physically leave a network such as schools and companies, they can still maintain connections with all the previous contacts. Although research claimed that SNSs can increase *bridging social capital*[42], it should be noted that the cost of social grooming is much lower than those done in the offline world. Given so many inter-connections between groups, different communities will gradually merge with each other and group structure will be effectively damaged. This is essentially the result
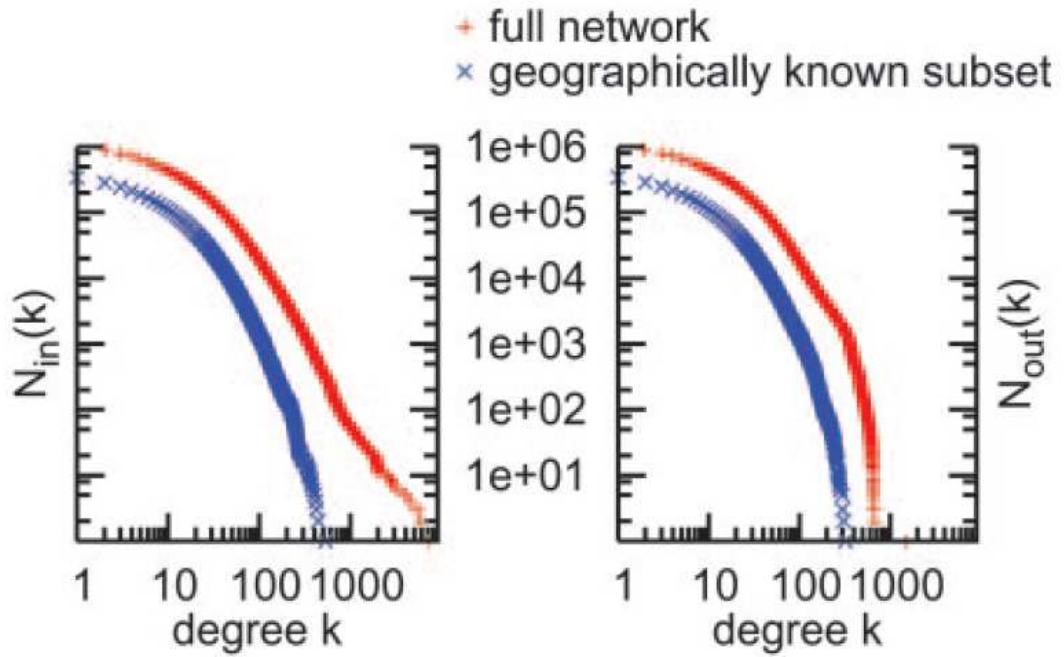
FIGURE 4.6: Degree distribution for livejournal.com. Taken from [88]
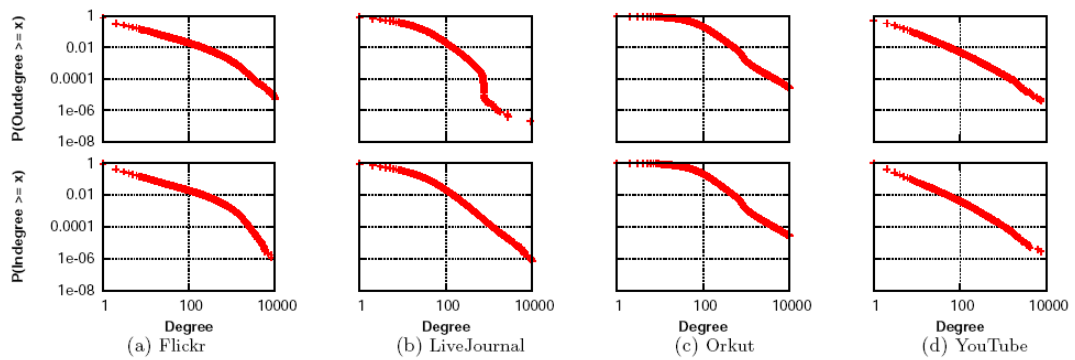
FIGURE 4.7: Degree distribution for livejournal.com. Taken from [93]

of dramatic increase of *bridging social capital* at near zero cost. Newman and Park have argued that group structure accounts for degree correlation in the network[99]. The degree correlation appears to be positive in social network but negative in most other networks such as information networks, technological networks and biological networks. Thus, positive degree correlation, which is also called assortative mixing or assortativity, can be seen as a unique characteristics of social network, in contrast to dissortative mixing in most other networks. The presence of assortativity signifies the likelihood for the complex network to be a social network. The assortativity of physics co-authorship, from example, is about 0.3[97].
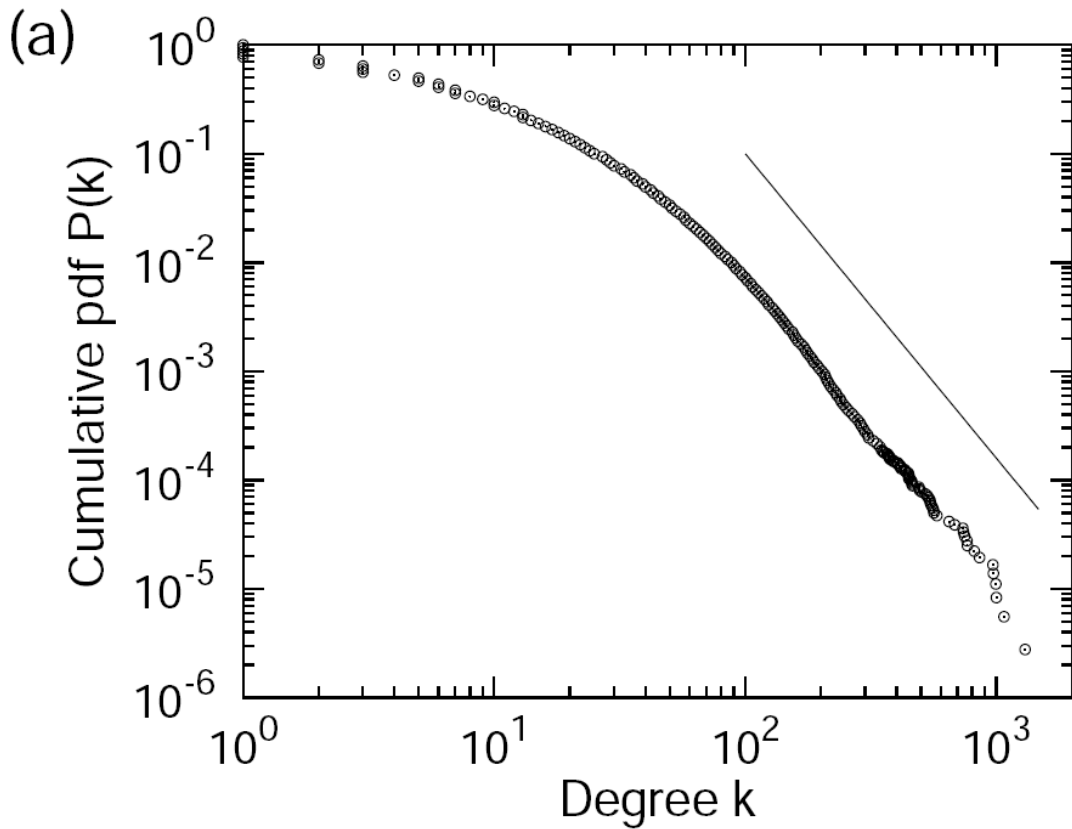
FIGURE 4.8: Degree distribution for mixi.com. Taken from [83]

However, given the impact of friendship inflation on group structure in the online social network, most established SNSs exhibit dissortative mixing or near-neutral degree correlation. Holme et al. measured the in-, out- and total degree correlations of *pussokram.com* and showed that all of these parameters are negative[63]. *pussokram.com* was a Swedish online dating website, which was launched in early 1999. The data was collected in February 2001. Mislove et al. calculated the assortativity social networks from both Livejournal and Orkut. The results indicate that the data is 0.179 for LiveJournal and 0.072 for Orkut, which are low or near neutral[93]. In particular, Livejournal was launched in 1999 and Orkut in 2004. Noted that the Orkut data, which contained information for about 3 million users, was collected between October and November, 2006. The Livejournal data was collected in December 2006. The assortativity of the Japanese site, Mixi, is about 0.125[83]. Noted that the data about Mixi, which was founded in 2004, was collected in March 2005. Ahn, Han and Kwak found that the assortativity is -0.13 for the Korean site Cyworld, 0.02 for MySpace and 0.31 for Orkut[4]. Noted that the Orkut data in this research, which contained information for about 100,000 users, was collected between June and September, 2006. The research suggests that the social network of Cyworld diverges significantly from real-world social network. This may be explained by the fact that Cyworld is three or four years older than MySpace and Orkut.
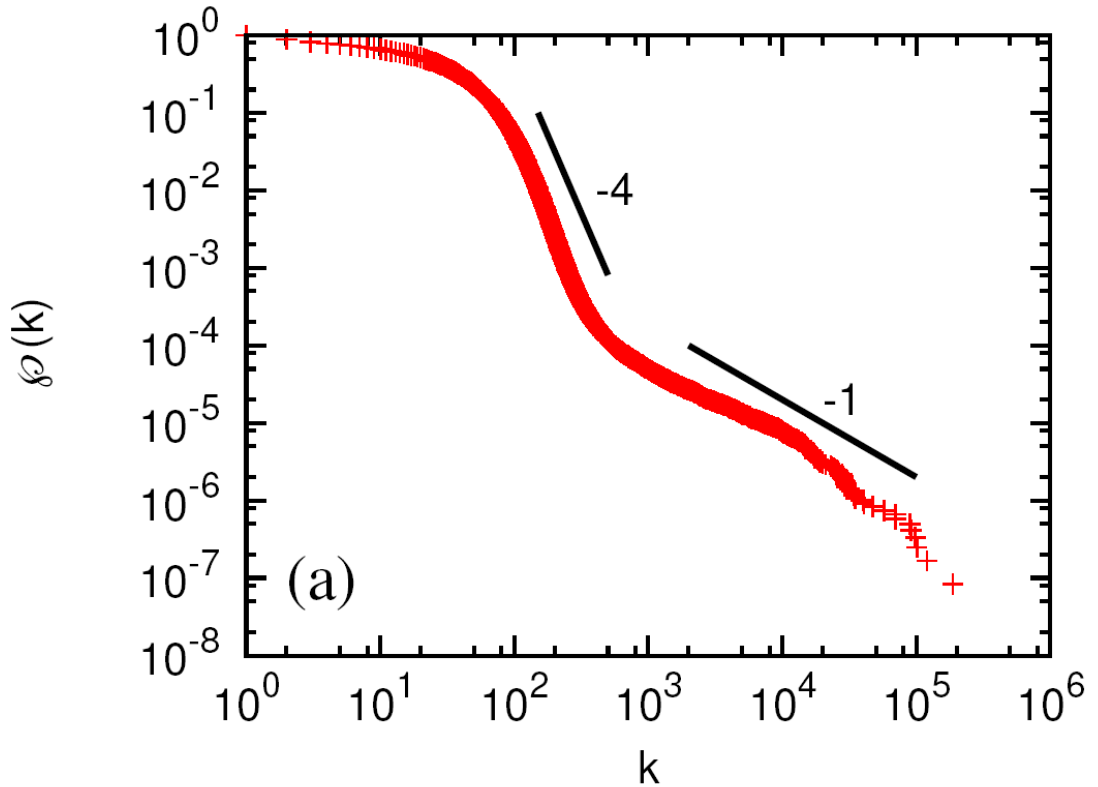
FIGURE 4.9: Degree distribution for cyworld.com. Taken from [4]

The data about Cyworld, which was launched in 2001, was collected in November 2005. In summary, many social network sites will exhibit a low value of assortativity. For the more established SNSs, the degree mixing can be below zero, which is the dissortative mixing pattern that is only found in non-social complex network.

4.3 The Network of the University of Southampton

Facebook was launched in the Harvard University in February, 2004. It quickly spread to other universities in the Ivy league. The site initially only allowed users to be registered with a university email address. In 2006, however, it started to open its registration to public. It is now among the top social network sites both in the US and among the world. Facebook introduces the concept of networks which refer to companies, organisations or cities any users belong to. Users can join up to two networks and may only change the network once every 3 months. Some of the networks, such as companies and universities, can only be joined with the proper university email addresses. For example, the network of the University of Southampton can only allow to be joined in with the email addresses ending with "soton.ac.uk". The network was established in September, 2006 and, as of this writing, has 24,518 members. More detailed information about Facebook and the
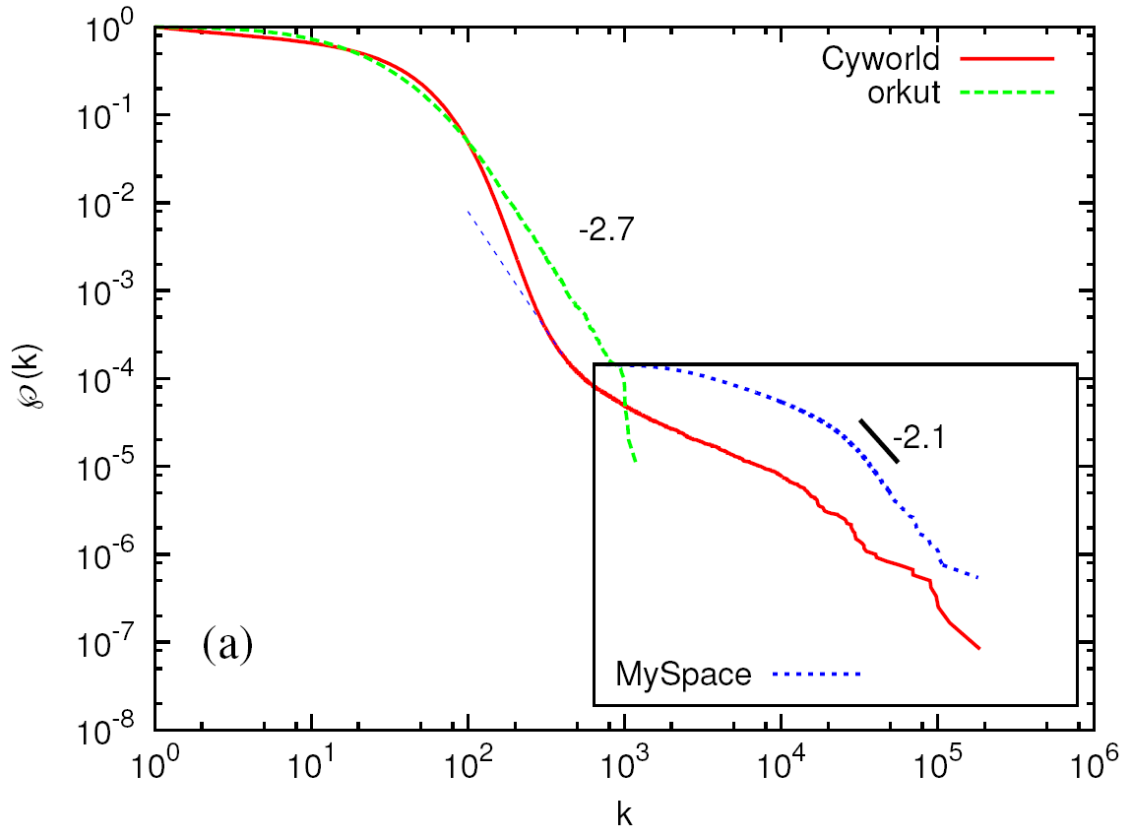
FIGURE 4.10: Degree distribution for myspace.com. Taken from [4]

demographics of the network of the University of Southampton can be found in Chapter 6.

Algorithm 1 Retrieve the social network of the University of Southampton on Facebook

---
Input: A Random University User on Facebook
Output: The Social Network of the University of Southampton S
ADAPTED-BREAD-FIRST(V)

1: Login on Facebook
2: Enqueue the root node V
3: **while** The queue is not empty **do**
4:     Dequeue a node
5:     Retrieve UID of V
6:     **for** All children of the node do **do**
7:         Enqueue the child node
8:     **end for**
9:     Sleep(10)
10: **end while**

We have attempted to contact with Facebook for access of the Data of the Facebook users in the network of the University, but did not hear any replies. Thus, we decided to crawl the data by imitating a normal user who is browsing the Facebook website. This technology may be called Web scrapping. The algorithm is shown as above. We

randomly select a node of the network. Any node directly linked to this node are then included in our data collection. The process is iterated until all nodes that can be reached from this node having be added to the final sample network. We managed to collected a sample of 15,005 in December, 2007, 19,604 in October, 2008 and 22,553 in April, 2009. It should be noted that users may change their privacy settings so that even other users in the same University network may not be able to access their list of friends. This problem, however, may sometimes circumvented by accessing their friends who are willing to list their friends. Some statistics of the data can be found in Table 4.11

	No. of Nodes	Mean Degree	Avg. Path Length	Clustering Coefficient	Assortativity	Diameter
December, 2007	15,005	63	3.08	0.16	0.32	11
October, 2008	19,604	67	3.12	0.16	0.20	11
April, 2009	22,553	73	3.13	0.16	0.34	9

FIGURE 4.11: Summary of data sets from the network of University of Southampton on Facebook

We begin our analysis of friendship inflation by looking at the growth of average number of friends on Facebook. A first look at these data reveals a steady growth of average number of friends of Facebook users in the network of the University of Southampton. The number increases from 63 in December, 2007, to 67 in October, 2008 and finally to 73 in April, 2009, as in the left graph in Figure 4.12. Giving Facebook's popularity in the University, it is not a surprising discovery that this number is increasing. We then investigate the initial network. This means we only look at the data set of 15,005 people in the 2008 and 2009 data collection. Theses users have been previously identified in our 2007 data set. The right graph in Figure 4.12 indicates that the number increases from 63 in December, 2007, to 66 in October, 2008 and finally to 72 in April, 2009. Thus, the growth of average number of friends is similar in the initial network to the growth network. We conclude that this growth does not only come from the early adopters of Facebook users but also from the users signing up in the following year, presumably the first-year university students.

Next, we compare the degree distributions of the three data sets. Figure 4.13 plots the complete graph of degree distributions in a log-log coordinate. The black line represents the degree distribution for the Sample 2007. The red line represents the degree distribution of the Sample 2008. The green line represents the degree distribution of the Sample 2009. All the sample networks exhibit a pattern of power-law degree distribution. However, in the scaling region of $50 \leq k \leq 500$, it shows that the probability p_k in both Sample 2008 and 2009 is bigger than that in Sample 2007, suggesting a monotonic increase of number of friends for vast majority of users, both active and less active. It also implies that the degree distribution is not scale-free, instead, it demonstrates a
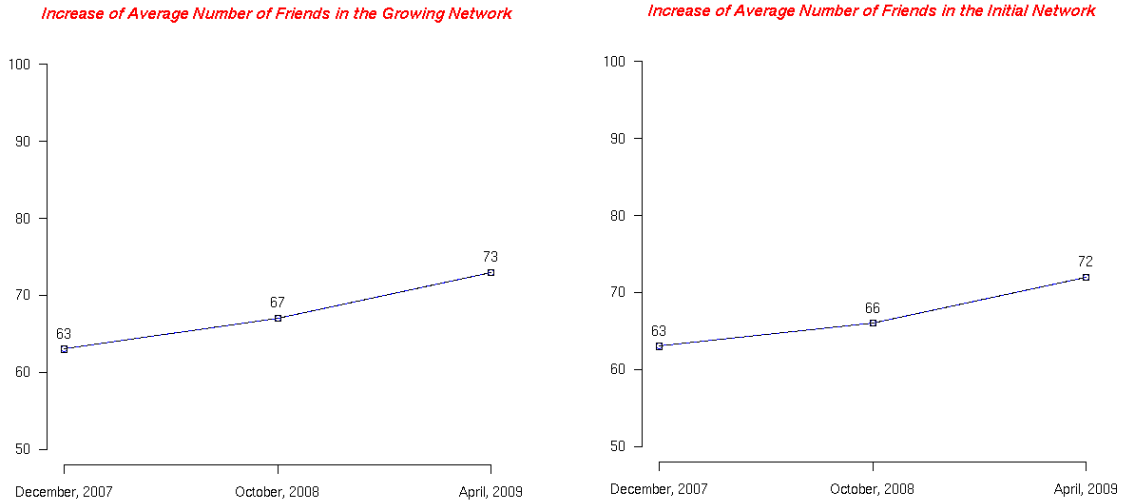
FIGURE 4.12: Steady growth of average number of friends of Facebook users in the network of the University of Southampton.

multi-scaling behaviour. In the region of $0 \leq k \leq 100$, the exponents α of all the samples are fairly similar, but beyond the region of $k=100$, this α becomes bigger for Sample 2008 and 2009. Besides, there is also a slight friendship inflation between the Sample 2008 and 2009.

A closer examinations of the degree distributions of all these three samples, as shown in Figure 4.14, Figure 4.15 and Figure 4.16, reveals the lack of clear cutoffs as discussed in the previous section. In particular, the degree distribution of Sample 2007 and 2008 will flatten beyond $k=500$. The degree distribution of Sample 2009 will exhibit similar patten when $k \geq 700$. The shortage of definite cutoffs implies that Facebook users care capable of befriending with more people at low cost by leveraging the technique of *static link*. To see how online social networks can empower the active users in the friend-making process, we select the people whose friends over 150. The number of 150, or Dunbar's number, is the supposed cognitive limit to the number of individuals with whom any one person can maintain stable social relationships. As shown in Figure 4.17, the number of people whose friends count over 150 is 1,273, or 8.5% of the sample population for Sample 2007. This increases to 1,869 in the second sample and 2,768 in the third one. There is even a bigger increase in the ratio of the number of active users and the whole sample population. It climbs to 9.5% in Sample 2008 and 12.3% in Sample 2009. The statistics clearly show that the degree distribution of highly active users do not obey the rule of scale-free behaviour. Active users will involve more on in friend making process.

The final metric we will investigate is assortativity. In our study, we focus on the connections between university members. Connections within the university represent a restricted relationships of the Facebook users. These relationships usually reflect the real connections as they stay in the same campus and city. As indicated in Table 4.11, the assortativity in all three samples is a relatively large positive value. However, we do
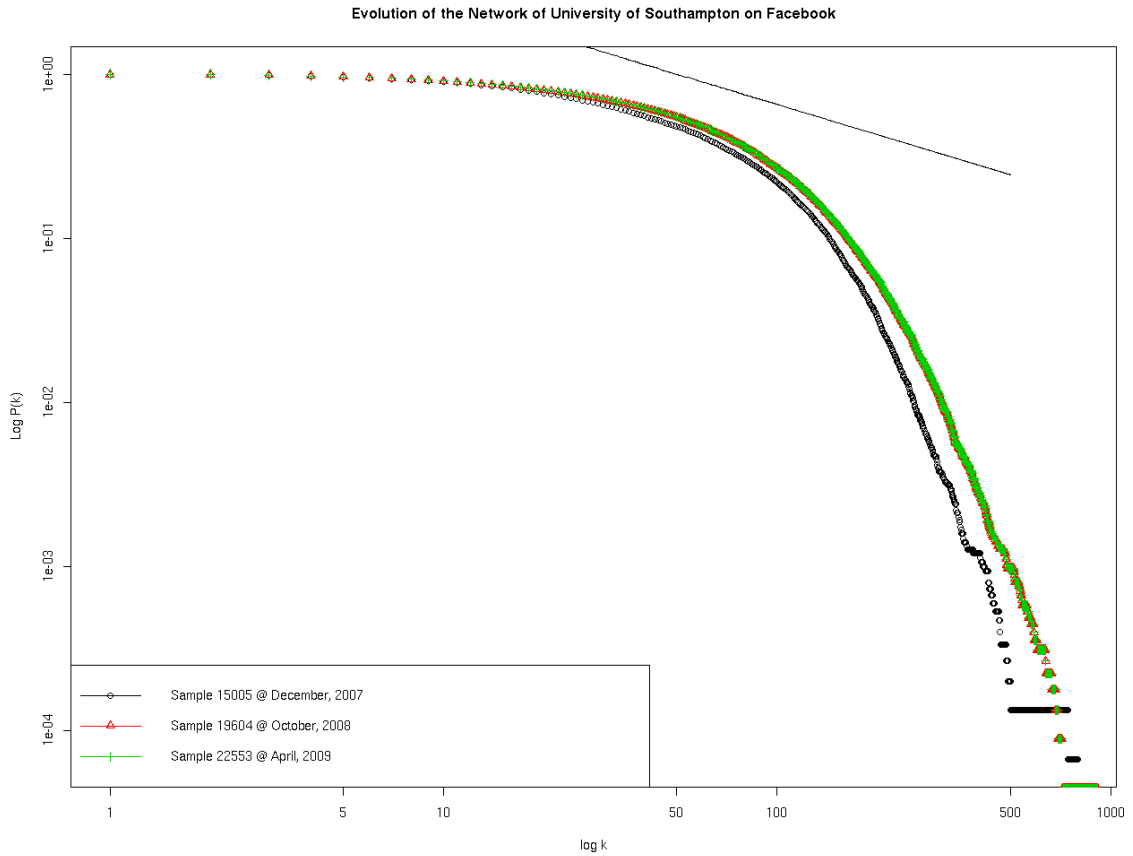
FIGURE 4.13: Comparison of Degree Distribution of the Three Data Sets.

observe the decline from the value of 0.32 in Sample 2007 to 0.2 in Sample 2008, which implies that the degree correlation moves from a bigger value to a smaller one. The change confirms prediction of the theory of friendship inflation.

4.4 Social Network Bubble

Social network sites allow users to browse others' social network by leveraging members' publicly articulated connections. The management of social capital is fundamental to social network sites. However, the problem of friendship inflation leads to an over representation of real-world social network. These can cause a lot of negative impacts on social network sites, most of which are not expected by system designers, who rarely consider the effect of social activities and behaviours of the users. It is increasingly difficult to distinguish the genuine connections in the social network, particularly the network of a more established social network site. More nodes in the social network appear to have hundreds to thousands of connections that can not be readily verified. Highly connected nodes and opinion leaders are increasingly undiscernible in the network. Users can no longer be held accountable for their behaviours due to the lack of peer pressure.

Sample 15005 @ December, 2007

Average Path Length l= 3.077417

Clustering Coefficient c= 0.1625199

Assortativity r= 0.3228468

Diameter d= 11
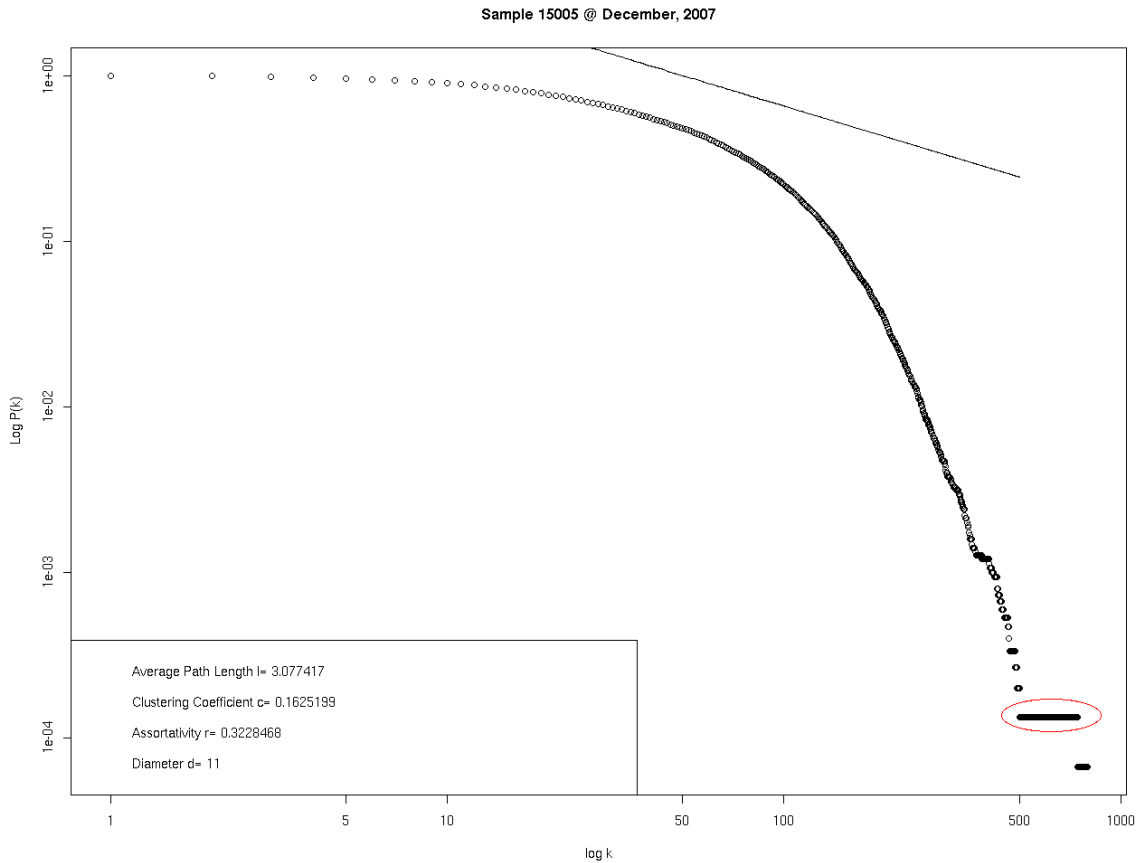
FIGURE 4.14: Topological Characteristics of Sample 2007

Given the mixture of the genuine connections and the fake ones, spammers and phishers can easily disguise themselves, spread junk information and collect users' data more aggressively. All graph algorithms that aim to take advantage of the social networks, such as Google's PageRank, Centrality Analysis and Community Structure Identification, will lose effectiveness because of the inflated number of edges. The technique of *static link* is also subject to users' manipulation and abuse. More connections lead to more information overload. This section will analyse these issues in detail.

4.4.1 Unreliable Connections

The merit of social network sites is that users publicise their private connections so that every individual can acquire new contacts via existing reliable connections. The practice is strongly encouraged and supported by most social network sites including Friendster and Facebook. The issue of friendship inflation will not affect users' ego-centric networks because they can recognise real connections in their own circles. The rule of 150, or Dunbar's number, indicates that for the connections that users are not able to recall from their memory, they are effectively not genuine connections at that moment. But when users attempt to traverse outside their own ego-centric networks,
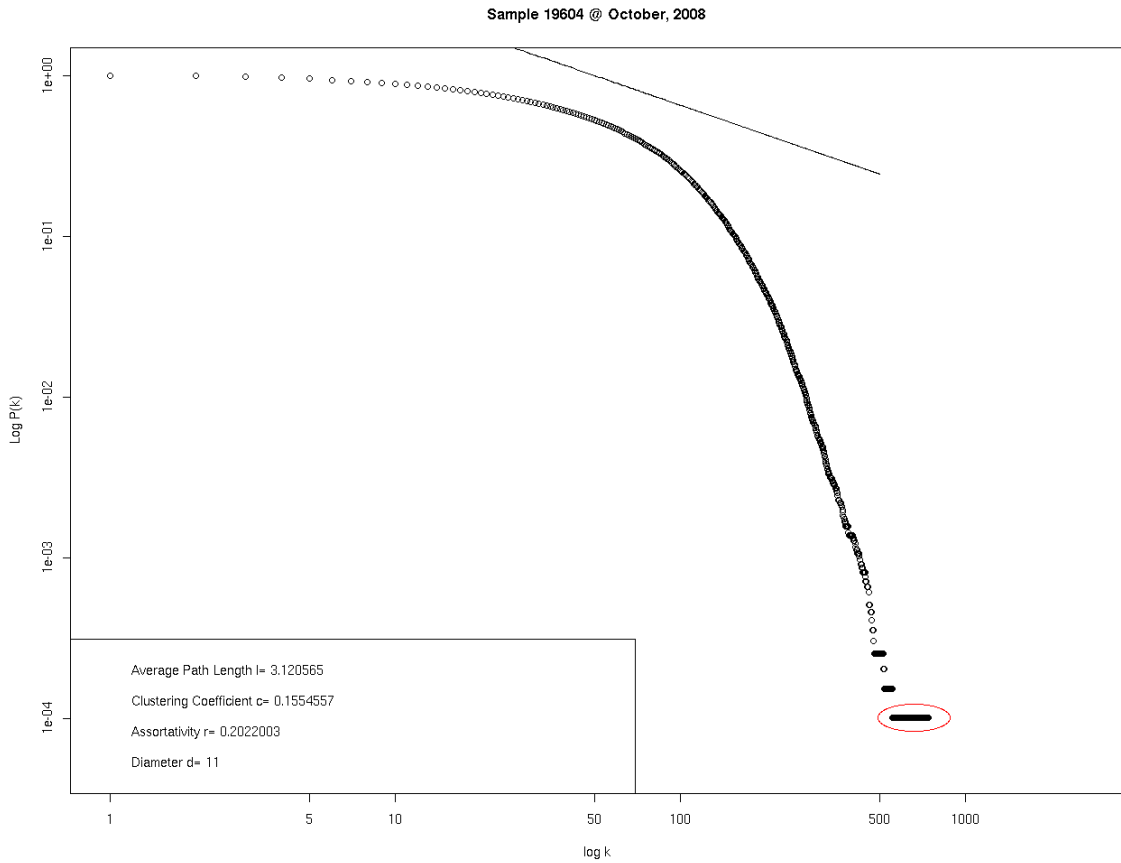
FIGURE 4.15: Topological Characteristics of Sample 2008

it does have a negative impact on the genuine structure of the social network. While users traverse their friends' social network, they will usually encounter a large number of so called "friends" whom they do not have any knowledge of. It is likely that they have never met these people in the offline world. To verify the connections outside ego-centric networks, they will then have to consult with their friends of direct connection in order to clarify their genuine friends who are meaningful to them. However, if users traverse the network further and go beyond two degrees, there is usually no such luck of consulting with people they can trust, for the connections are so remote that it is hardly to clarify the relationships with them. As the network grows over time, they are increasingly cautious to approach people out of their ego-centric network. More careful observations have to be made to identify reliable connections. It is no longer convenient to trace the credibility of acquaintances. Their friends can no longer serve as meaningful connectors and recommenders. Friendship inflation makes the navigation of online social network becomes increasingly like navigating in a place of strangers. Rather than saving one's time and effort for establishing trustful relationships, the tedious navigation may cost more time than the direct consultation with one's friends.

Boyd pointed out that on *Friendster*, when traversing the network, there is no way to determine what metric is used or what the role or weight of the relationship is. The
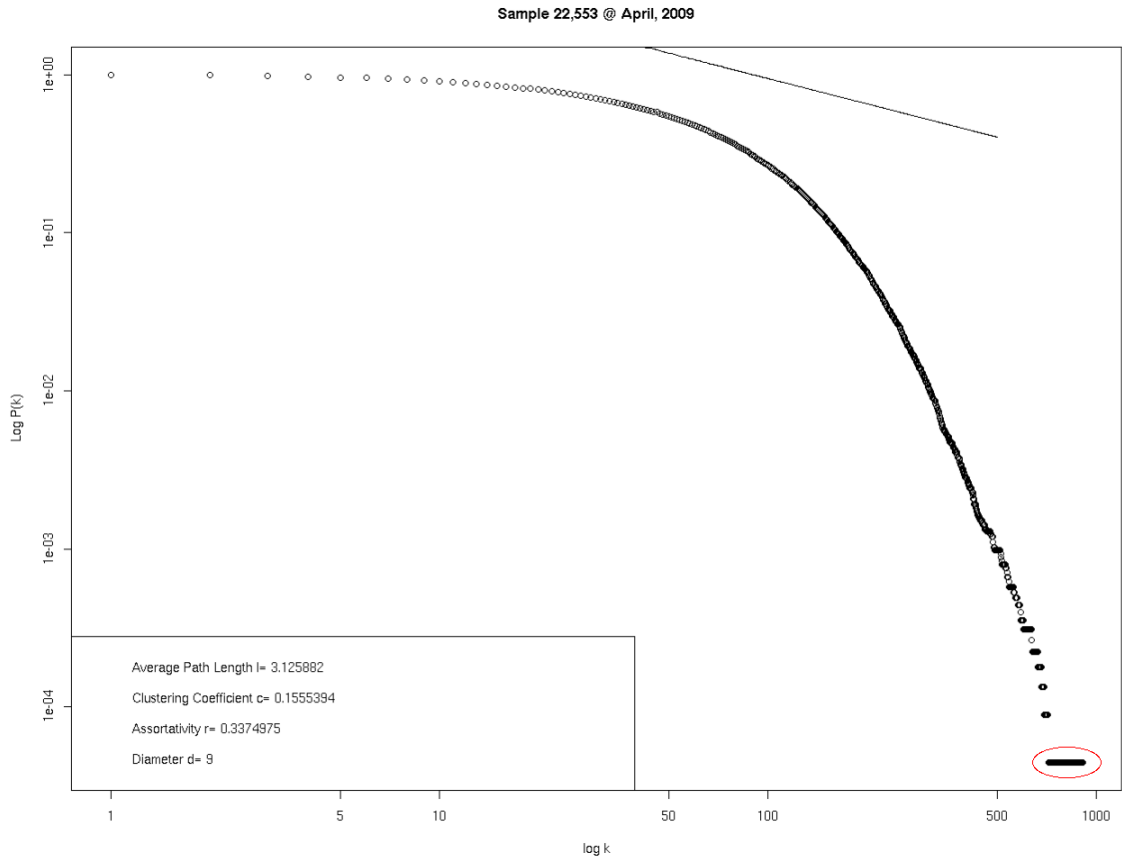
FIGURE 4.16: Topological Characteristics of Sample 2008

phenomenon repeats on both MySpace and Facebook. For example, on MySpace, Tom is the first person that a new user will add to his or her friend list when he or she logs on the website for the first time. In theory, MySpace Tom can have a number of friends roughly equivalent to the total size of the social network. As of this writing, MySpace Tom has about 238,660,532 friends, as shown in Figure 4.18. The huge number of friends is a good example on the difficulty of distinguishing reliable connections from strangers and acquaintances. Once the network goes beyond the point of real-world connections, the problem starts to emerge, until users get bored with the site and leave.

4.4.2 Undiscernible Hubs

The highly connected nodes in the social network are usually opinion leaders and centres for information sharing and dissemination. In the offline world, these people can be singled out by how many connections they maintain. Other factors like how popular they are in a given community can also indicate that they are the opinion leaders of that particular community. Ordinary users may pay more attention to this particular group of people (by subscribing to their blogs, for example) if they are interested in what these people say and publish. Because of friendship inflation, some low-key users
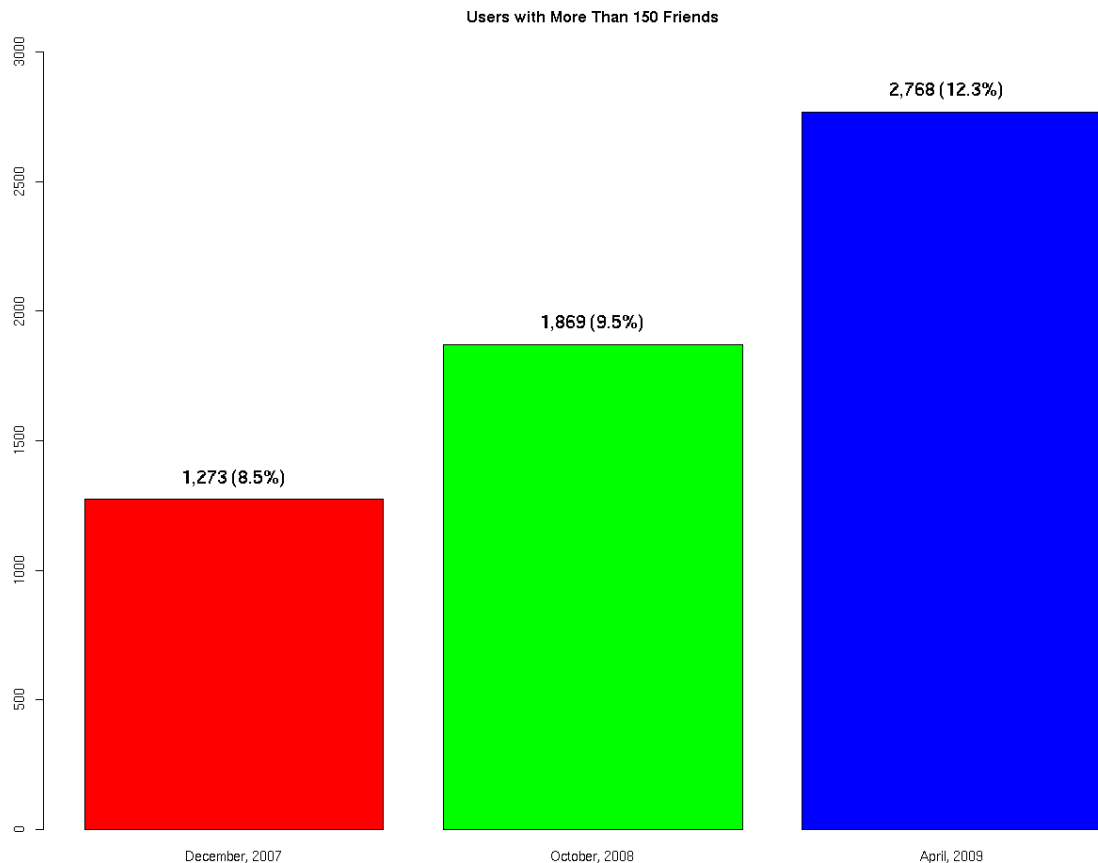
FIGURE 4.17: Increase of Average Number of Friends of Active Users.

can appear to have as many connections as those genuine high-profile users. These users can be friendship collectors, who may attempt to acquire contacts in an aggressive manner. They can be fakesters and fraudsters, which are created by ordinary users for the sake of connecting to other users who share similar interests and social and cultural backgrounds. They can be spammers who relentlessly gather contacts in order to send out junk messages and commercial advertisements. They can be phishers who seek to acquire sensitive information by deliberately accumulating the connections with other users. The emerge of non-hubs that appear to share similar number of connections of real hubs increases the cost of looking for a real one. To verify a real hub, a user may have to go into details on what they have published, how they communicate with others, what photos and videos they have uploaded on the site, etc. The combination of real hubs and fake ones also affects system designers who are targeting opinion leaders for viral marketing. Influential and prestige members who can accelerate the dissemination of information can no longer be readily recognised. There are many network-based algorithms, such as Google's FriendRank, claimed to assist targeting advertising by identifying the information centres of the social network. Unfortunately, they fail to understand the issue of friendship inflation. As a result, the hubs and centres that these
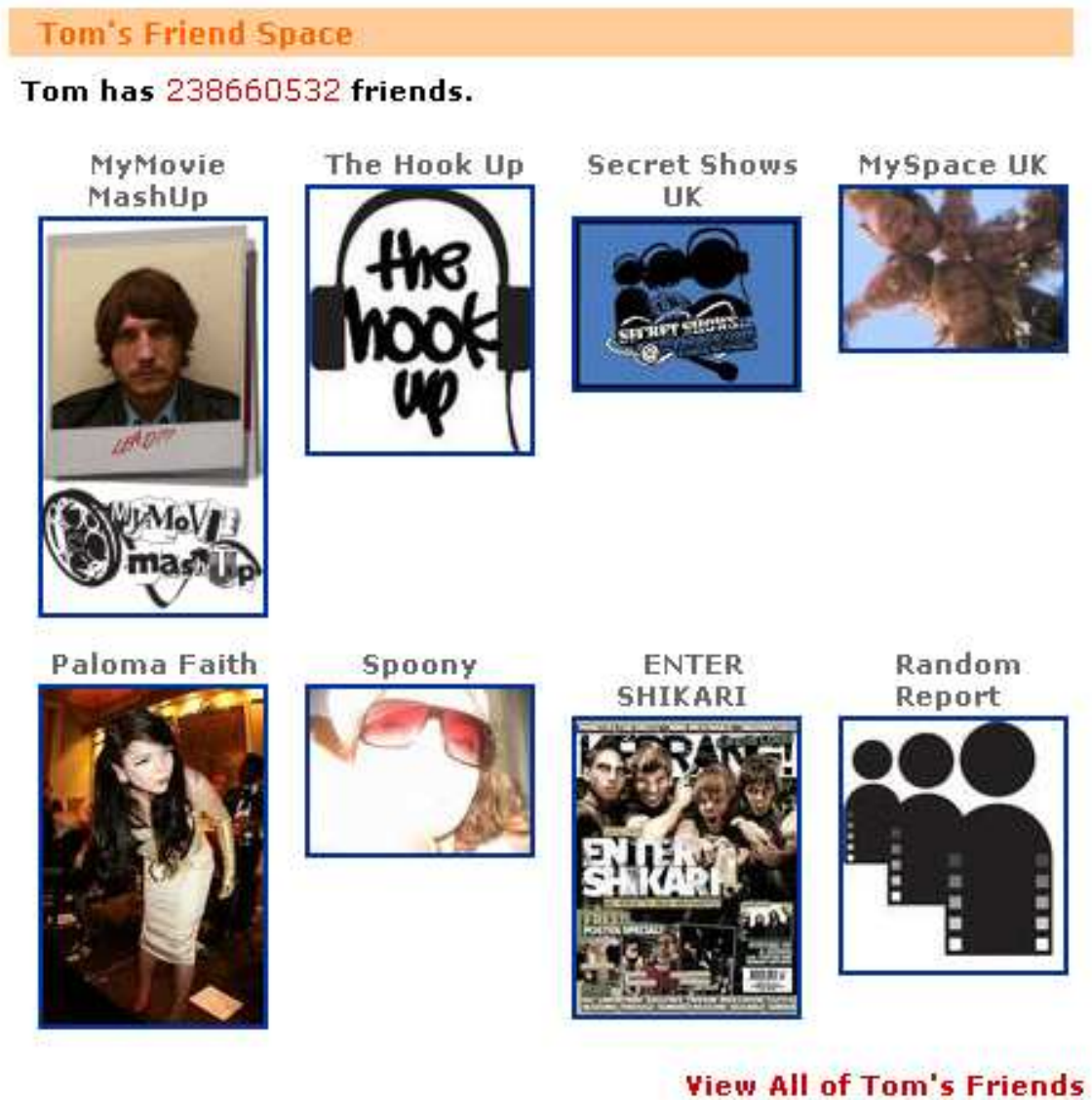
FIGURE 4.18: Tom's Friends on MySpace

algorithms have identified are essentially those appear to merely have a sheer number of online contacts.

The cases of *Friendster whore* and *Facebook whales* are good examples of undiscernible hubs. *Facebook whales* are the popular users who actually have a lot of genuine connections. These can be bloggers, journalists and celebrities. On the other hand, *Friendster whore* are users who deliberately collect friends to boost their popularity. Because of the *static link*, it is legitimate to acquire as many connections as possible in a short period. There is no way to tell the difference between *Friendster whore* and *Facebook whales* without actually scrutinising their activities and behaviours on the social network sites. Ironically, the *Facebook whales* are usually so busy in coping with their friends that they may not spend sufficient time in decorating their profiles and upload materials on the

website. In contrast, the *Friendster whore* will have plenty of time to decorate their profiles, making them much appealing to ordinary users. As a result, the close examination of their activities and behaviours may end up with the opposite conclusion.

4.4.3   Lack of Peer Pressure

On the Internet and the Web, users can publish information and still remain anonymous. It is difficult to hold them accountable if they are spreading rumours and unconfirmed stories. On the contrary, social network sites are supposed to be the place where individuals can be identified and held responsible for their online behaviours and activities. The trustfulness of any individual can not only be assessed by the information of their profiles but also the people directly connecting to them. The people are supposed to be genuine friends of the users. The more friends one has, the less likely that he or she will take the risk to involve with behaviours and activities that will damage their reputation and credibility. However, in the hyperfriendship social network, the number of friends have been inflated and can no longer be used to determine one's identity. Many users appear to have a large number of connections without revealing their true identities. As a result, it is difficult to apply peer pressure to these people because their peers are simply not their real life friends.

The problem of lack of peer pressure can be seen on the materials that have been published and uploaded on social network sites. In MySpace, for instance, it is not unusual to see explicit materials such as pornography on some users' profiles, which, however, are published by users with hundreds to thousands of contacts. It is unimaginable that they will attempt to publish the same materials if these materials are to be viewed by their genuine friends in the real world. However, on social network sites, those "friends" can simply be acquaintances and therefore the users who publish the explicit materials can not be collectively held responsible for their activities. The same phenomenon appears in online transaction. It is reported that users who involve in buying and selling have more satisfaction than traditional websites, such as Amazon and eBay. This is largely based on a social network where the connections are trustful. Malicious sellers and buyers can be tracked down through a chain of reliable connections if they exhibit bad behaviours and activities during the course of transaction. Bad reputation will spread across the whole social network via word-of-mouth. Fears of bad reputation will prevent people who involve with the transactions from making risking decisions. However, if the contacts of these sellers and buyers are no longer genuine friends, no peer pressure can be applied to them. They do not have to worry about their reputations in their ego-centric networks. The transactions will become less and less trustful as the network grows.

4.4.4   Spamming and Phishing

Friendship inflation is a good disguise for spammers and phishers who seek to establish hundreds of thousands of connections in a short period without much effort. With this information in hand, spammers can spread junk information across the network or send out commercial advertisements to users. The spamming on social network sites can even be worse than traditional computer-mediated communication software. Email filters, for example, stop the junk mails by assuming that unsolicited messages about commercial products such as medicines and fake university degrees are universally undesirable. This is not completely true on social network sites, because messages can usually be sent to people who have acquired mutual connections. SNSs assume that people who have already befriended with each other will send and accept all types of information. Therefore, the social network system will not attempt to detect any spams in these established channels. Spammers do not have to circumvent around the spamming filters in order to broadcast the unsolicited messages across the social network. Users who frequently receive junk messages from the spammer contacts may decide to end the connections with them. However, as most social network sites are open to public registration, including Facebook and Okurt, which were previously only limited to university users, spammers can change their registered email addresses and establish a new stock of thousands of connections in a short period with only a handful of clicks. Many social network sites do have mechanism in detecting irregular activities and malicious behaviours, they cannot, however, understand the difference between the friend request from a spammer and that from a normal user who wants to make genuine friends on the site. Phishers who seek to acquire sensitive information from SNS users can also benefit from friendship inflation. They are equally aggressive in harvesting profile data from other users. Given the problem of friendship inflation, phishers do not even spend a lot of efforts in social engineering in order to obtain users' private data. They only need to befriend with the contacts of the targeted users by making some superficial connections. With a few hundred of real contacts, phishers can win the trust of the targeted users and have access to their personal data.

Social network sites are quite vulnerable to these attacks. They usually have to resort to legislation and law enforcement. MySpace, for instance, recently succeeded to sue a so-called spam king for allegedly using compromised user accounts to send millions unsolicited advertisements touting ring-tones, polo shirts and many other things. But for many other spammers, they can only detect them case-by-case. There is a lack of generic technique in dealing with spamming and phishing in the online social network. The fundamental weakness of contemporary social network sites is the use of *static link*, which incubates a large amount of casual connections.

4.4.5   Inaccuracy of Network-based Algorithms

Many network algorithms aim to capitalise the rich resources of connections in the social network. These include algorithms based on different kinds of centrality and prestigious analysis such as degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. Besides, there are also complex network algorithms such as walking in the network and community structure analysis[86]. Google, for instance, has developed a method for displaying advertisements to members of a network based on its PageRank alogrithm[1]. Other research area focused on the community evolution by studying the edge change of the network. All of these algorithms take for granted that the topology of online social network is the same as that of real-world social network. Unfortunately, as we analysed earlier, this network is a hyperfriendship network where the number of connections are always increasing. The resulting network is a super set of real-world social network. The percentage of highly connected hubs in the hyperfriendship network is much less than that in the real network. The mixing pattern in the hyperfriendship network appears to be very small or below zero, compared against a larger value of the real network. Many nodes which score very low in real-world social network will have much higher centrality in the hyperfriendship social network. Community structures will be increasingly vague due to the crossings between various groups, – the permanent connections that have been established through a handful of clicks. Without the character of *preferential attachment*, the diameter of the network will grow bigger, making the average short path become longer. The inter-connections between different groups will also play a negative impact on the accuracy of the calculation of clustering coefficients. This will cause inaccuracy and even fatal error to the algorithms, on which a significant part of patents are based.

Consider the case of Google's PageRank. Because it utilises the link structure of hypertext, webmasters can take advantage of the algorithm by inter-linking their websites with other webmasters. This is called link farm. Google actively penalises the link farm because they will inflate the score of PageRank. In social network sites, however, users are perfectly legitimate to boost their "PageRank" by adding as many friends as possible. They can not be punished on the ground that they have too many fake friends, as in the case of link farm on the Web. With the convenience of befriending, some even attempt to manipulate their popularity index by collecting more contacts. As a result, algorithms such as FriendRank, an algorithm developed by Google for displaying advertisements to members of a network by using eigenvector centrality, will fall into the trap of the link farm in hyperfriendship social network.

---

[1]http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FP

4.4.6   Information Overload

The increase of connections leads to the increase of information channels. Besides synchronous channels such as instant messaging and asynchronous channels such as private messaging and public wall posting, there are a whole range of indirect communication methods such as status updates, commenting on blogs, photos and videos and pokes. As many social network sites do not yet provide adequate tools for fine-grained management of the information sent from their friends, users can easily end up with a large amount of messages and information from their direct contacts. We have previously argued that social network is particularly effective in information sharing and dissemination. Information can propagate in the social network much faster than other type of networks. The hyperfriendship social network, which is a superset of real-world social network, can only exacerbate the problem by introducing more short-cut paths between users. For instance, in physical world, when one leaves schools or companies, he or she will generally communicate less with the previous schoolmates or colleagues. The connections will thus gradually decay. On the other hand, he or she will establish new connections in the new school and company. The rule of 150, or Dunbar's number, suggests that connections can not keep increasing but there is always a cap on human beings' cognitive memory. However, on social network sites, human being's shortcoming is compensated by the hard disk storage, which can maintain huge amount of temporal contacts. The connections previously only appear in one's private address book, if visible in the social network, will quickly form an information network that can channel enormous amount of information very quickly. As a result, the idea of "communicate with anyone anywhere" is quickly translated to "flood anyone anywhere".

Facebook opens its platform and allows third parties to develop applications on it. Some users are happy to use these applications while others remain indifferent. However, if applications are accepted and used by a user, Facebook will send an invitation to his or her friends, unless they have changed their privacy settings to reject such notification messages. As a result, even the users who want to stay away from those applications will still receive a lot of information from their peers who are using these applications. Another case will be news feeds, which report to a user about his or her friends activities. With friendship inflation, many news feeds actually come from the acquaintances who may have never met the users before. Given many connections are weak ties that are induced by friendship inflation, such information channels are subject to abuse such as spamming and phishing. But even we exclude the case of spammers and phishers and only consider ordinary users who are willing to share news and information. The information overload is still quite obvious.

4.5 Boom and Bust of YASNS

As of this writing, some statistics show that Facebook just overtook MySpace for registered users as the leading social network site. But when it comes back to the year 2004, it was Friendster that was once on the lead. There are so many social network sties available today that a new one is sometimes called Yet Another Social Network Site (YASNS). If history is any indication, then YASNS comes and goes. It may well be argued that the rise and fall of social network sites is mainly due to the improvement of technology over time. For example, when SixDegrees launched in 1997, digital cameras were not as popular as today and therefore photo and video uploading was non-existent. It is also true that at that time, many Web technologies such as JavaScript, PHP and Flash animations were not mature yet. Internet connections were also much slower than today's broadband connection.

However, it is equally true that given the huge amount of users the established social network sites already have, they enjoy a competitive advantage to the late comers, as manifested in the network effect. There is no reason that users prefer a new site to one where most of their friends have already registered with. Besides, it is not difficult to copy the improvement of new sites. In fact, major social network sites such as Facebook, MySpace and Bebo all share similar functionalities such as profiles, blogging, private messaging, photo and video uploading, discussion groups, etc. Therefore, we do not consider these issues are fundamental to the rise and fall of social network sites. Instead, we argue that friendship inflation, which devalues the whole network, is the main reason for the decline of social networks.

4.5.1 Cases of Friendster, MySpace and Facebook

Friendster launched in 2002 and rapidly amassed over 5 million registered accounts by January 2004. However, at that point, Friendster had already witnessed massive friendship inflation. Over time, members have accrued a large amount of connections yet there were no metrics to indicate the weight of the connections. The connections are typically binary: friends or non-friends. It is so convenient to befriend on Friendster that some users create fake profiles in order to attract other users who share similar interests and social and cultural background. These were called Fakesters. Friendster considered that fakesters devalued the social network and therefore decided to remove them from the site. The massive removal of fakesters without consulting with users who created them have annoyed a lot of users. The so-called Fakester genocide exacerbated the situation and had driven many users to MySpace, which was then a new competitor to Friendster. It should be noted that today Friendster is still popular in Philippine and Southeast Asia. This suggests that technology was not the major factor to be blamed for the decline of Friendster in the US. In fact, when MySpace emerged, Friendster had just made an

improvement on their system by adding more servers and changing the programming language from cumbersome Java to light-weight PHP[104]. The same argument applies to Facebook when it just arrived in the arena of social network competition. But eventually Facebook outweighted MySpace, which had previously outweighted Friendster. Today, people have already been talking about Facebook fatigue, given that the number of registered users dropped slightly. Now with the rising of Twitter, Facebook may well be another victim of friendship inflation.

One of main attractions of social network sites is to make new friends by leveraging the existing connections which are assumed to be reliable. Friendship inflation suggests that users will have more contacts than they actually have in the real world. In a social network with nodes densely connected with each other, it is very difficult to make meaningful connections because the substantial cost of discerning the real connections. The sites will eventually lose their original advantages. SNSs may gradually re-position themselves in the competition with the new sites. For example, MySpace looks increasingly similar to a media site by providing videos and music. Facebook looks increasingly similar to a communication tool by providing web-based instant messaging and twitter-like activity updates. When they fail to reflect the evolution of users' social network and capture the real network, if there are new alternatives, users may just leave the old site and switch to the new network for the search of genuine connections. Here, the balance point is the effort to distinguish the genuine contacts from strangers and acquaintances in the old established social network site, versus the effort to invite friends to the new site. Because of the *static link*, it is always easier to establish connections on the site. Therefore, as the old site becomes more crowded and trustless, the new site will be more attractive.

4.5.2 summary

Hyperfriendship social network provides no mechanism to verify the connections between users. There is not any rule by which users must adhere to establish new connections. Users can make new connections without much cost. This leads to rampant increase of the number of friend connections. The rapid growth of hyperfriendship social network collapses the context and social environment where users interact with each others. The increasing of the weak yet persistent connections bring a whole range of social implications and ramifications, complicating the issues of fakesters, privacy concerns, multiple personas, spamming and phishing. Social network sites seek to tackle the issues by using different technologies, human interventions and even resorting to law, but with *static link* as the backbone connecting method of their network, most SNSs are fighting a losing battle on balancing between publicity and privacy. The hyperfriendship network can be saturated but users' real-world network is still evolving. When an alternative social network emerges, users who are fed up with the old one simply switch

to the alternative. The critical issue for most social network sites is that they attempt to configure users by the technologies they have developed. The technologies include *static link* and contact categorisations. System designers hope that users will use these technologies and use them in a way that conform to their intention, which, according to Friendster and Facebook, is to encourage genuine identity and connections. The vision is shared by most social network sites. However, when it comes to friendship collectors and fakesters and fraudsters, system designers simply ignore the creativities of users and brutally configure the users according to their original intention by limiting or removing the unwanted activities.

4.6 Discussion

In this chapter, we present statistical evidence in macro level that supports friendship inflation in most social network sites. Two outstanding statistical properties are no definite cutoff and dissortative mixing pattern. The theory of friendship inflation is supported by our nearly three-year observation on the Facebook users in the network of the University of Southampton. Then, we discuss the issues incurred from friendship inflation. The problems include unreliable connections, undiscernible hubs, lack of peer pressure, spamming and phishing, inaccuracy of network algorithms and information overload. We argue that friendship inflation will eventually lead to the decline of social network sites. To support the argument, we cite the case of rise and fall of Friendster, MySpace and Facebook. In next chapter, we will present the algorithm of *ActiveLink*, which aims to solve the problem of friendship inflation by identifying meaningful social connections.

Chapter 5

ActiveLink: Identify Meaningful Social Connections

5.1 Introduction

Most issues confronting social network sites come from the fact that they are modelling people's dynamic ever-changing real-world connections in a static framework. The static model holds an implicitly stationary view of relationship formation in which connections, once formed, were permanent – thus entailing zero maintenance cost[25]. The static model damages the properties and topology of real-world social network. It fails to reflect the evolution of the network. Unfortunately, there is little academic research being done to address the fundamental issues of the static system, despite more and more commercial and experimental social network sites available. We propose the algorithm of *ActiveLink*, a communication-based method that aims to identify the genuine connections.

5.2 Evolving Social Network

Real-world social network is an evolving social network. People and their social connections are constantly changing. The existence of a network of connections is not a natural given, constituted once and for all by an initial act of institution. Instead, it is the product of an endless effort of material and information exchange which presupposes and produces mutual knowledge and recognition[15]. People acquire new contacts as they advance in their schoolings and careers. Old connections may gradually decay over time. When people become elder and less engaged in social activities, they tend to lose previous connections and attract fewer new ones. Unfortunately, the vast majority of contemporary social network sites, which mainly employ the technique of *static*

link, fail to capture the evolution of the network. When the network size inflates, it will usually be brutally re-configured by cutting down the number of connections and removing the profiles that appear to be fake. Some will force the users to think carefully when adding new connections by imposing an upper limit to the maximum number of friends, but usually encounter massive rebellion from the users and subsequently drop the cap. In some cases, social network sites build a second tier of top friend network to mitigate the issues, only to find more discrepancies of the connections. The fundamental weakness of many social networks is the static framework of an evolving network, which turns out to be a failure, leading to the bust of Yet-Another-Social-Network-Site.

However, as we stop using the *static link* and start to model the dynamic social network, it opens a door of uncertainty. A lot of questions will emerge from defining people's connections. What are dynamic and meaningful connections? How often should people interact with each other in order to be counted as "active"? How many connections can people maintain? Once these connections are established, will they decay in the future? And if so, how long can it last? What maintenance does it take to keep the connection alive? Does the rule apply universally to all the people? To answer these questions, we introduce the idea of *ActiveLink* in the following section.

5.3 ActiveLink

Active links refer to the connections between users who often exchange and share information. The methods for exchanging messages include both direct communications such as private messaging and instant messaging and indirect communications such as public wall posting, blog commenting, photo and video commenting and gift exchanging. Instead of assuming a zero-cost establishment of connection, it levies certain amount of communication "tax" to maintain the connection. The idea will be translated into the practice that the system will no longer employs the *static link* that takes a few clicks to befriend with one another, instead, it will look at how users communicate with others whom they have added as friends and only the presence of continuous communication signals connection. Many social network systems which recognise the weakness of the social networking technique of *static link* may attempt to devise a new algorithm for social network connection based on users' behaviours and activities. However, while reciprocity is at the heart of these algorithms, they rarely consider the role of already-established social capital in determining the number of connections each user can acquire. The algorithm of *ActiveLink* is designed to be consistent with some topological features as found in the social network, such as *Preferential Attachment* and *Assortativity*. It will also take into account the factors of ageing and cognitive limit of human beings' brain. To illustrate the model, we compare the network of active connections with representative democracy model. At the end of the section, a algorithm is given to illustrate our ideas introduced in this section.

5.3.1 Continuous Reciprocity

There are many online activities that can signal the existence of genuine connections. For example, a user's online actions towards his/her friends and monetary transaction may both signal the genuine connections. However, there are several problems with these methods. First, the unilateral action fails to capture the mutual recognition between any two users. The method can easily be abused by users. Monetary transaction may indeed reflect genuine connection. However, it seems the transaction only happens between small percentage of total members as social network sites mainly facilitate information flow rather than cash flow. Thus, the transaction method will under-represent the real social network.

As mentioned earlier, reliable connection is an endless effort of material and information exchange, we therefore propose that the active connection can be based on direct communication such as private messaging and indirect communication such as wall posting, mutual blog commenting, photo and video commenting and gifts exchanging. Signalling theory states that each agent has qualities that they wish to communicate. The length, frequency, and content of public comments and other communication signal the strength and context of a relationship and do so with greater nuance[33]. To simplify the model, we only utilise the frequency of communication. The choice is partially intuitive, and partially due to some available research which suggests that the more people communicate, the closer they are[59].

For a connection to be *active*, the frequency of communication must be no less than a certain threshold, which we denote as *f*. It should be noted that whereas we do not specify the forms of communication, they contain both public and private information exchanging. The public method generally includes wall posting, blog commenting, photo and video commenting and gift exchanging. The private method generally includes private messaging and instant messaging. We consider that both are equally important for the connection, though it may well be the case that people who communicate via private messaging may well have better relationship than those via photo mutual commenting. The same weight attached to both communication methods can also benefit the users in that it allows users to communicate in their convenience, without the bias to prefer one over another simply because the method can contribute more to their ego-centric network.

5.3.2 Contact Cap

For any given community, there is an average number of regular contacts each person can maintain. Note that the average number refers to the median rather than the mean, since there exists a small amount of people who can manage disproportional large number of connections. The British anthropologist Robin Dunbar proved that the upper limit is

about 150[37], with evidence found not only in ancient villages and tribal groups but also in modern armies and companies. However, the connection cap in real-world social network is usually smaller than that in online social network. For instance, some research suggests the number on Facebook stands at about 250[53]. Many users can maintain the number of contacts well above 150. This is because social network sites can assist users to manage their social networks by providing tools for contact storage and friendship management. It is much more convenient to groom in online social network than other offline methods. Given the social network sites as a social capital management tool, we argue that the connection cap C can be greater than the Dunbar's number. The median of connections m should be any number less than C.

5.3.3 Connection Decays

We establish new connections as we communicate with acquaintances. The more interactions take place, the more durable the connections will likely be. The durability of the connections is supposed to fade away gradually if we do not keep in touch with our friends. We represent this in our model by giving each a **connection strength** S. When person u interacts with person v , the strength $S(uv)$ of the connection between them is set to be 1. Then as time passes, the strength S decays exponentially if they do not exchange information[68]:

$$S(uv) = e^{-k\Delta t}$$

Where k is an adjustable parameter of the model. For our convenience, it is set to be 0.001 in our case. Figure 5.1 shows the change of **connection strength** over time.

If they communicate again, $S(uv)$ is set back to be 1.

For our convenience, we suggest the period for connection expiration D is 50 days. After 50 days, v becomes an *inactive contact* of u.

Research in experimental psychology has demonstrated that there is a decline in memory retention over time, commonly known as *Forgetting Curve*. The formula describing the forgetting process is similar to the one we employed to describe the strength of connection[40]. The coincident reflects the decay of old friendships in our real life as we move to a new stage and explains the fact that we would spend time and effort on maintaining the existing relationships which we do cherish. The application of active contact can effectively exclude casual acquaintances as we pointed out in Chapter 3. We do not communicate with acquaintances as frequently as we do with close friends. But we would keep these people in our contact lists due to the weak tie assumption. In the future, if we communicate with them for some reason, then they will be "activated" and become our active contacts. Therefore, the concept of active contact and strength of
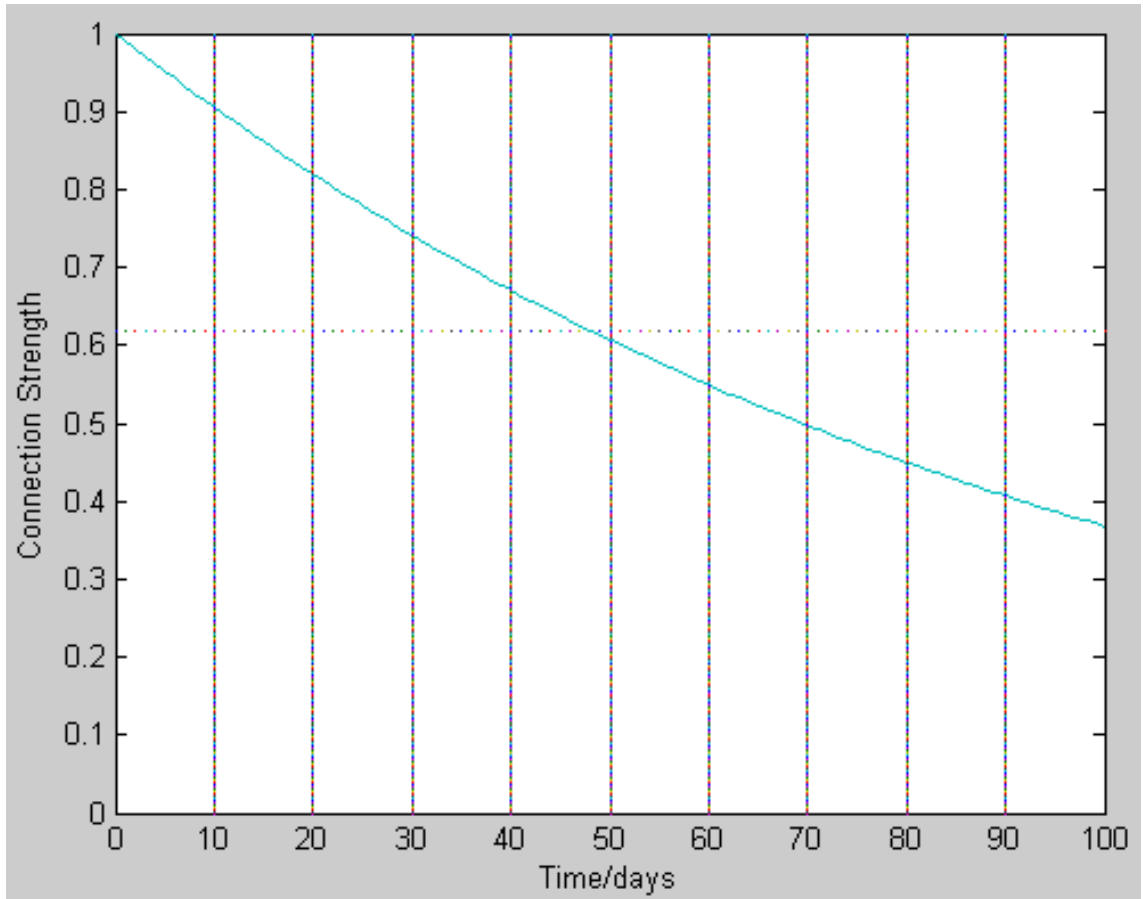
FIGURE 5.1: The Connection Strength - Time Diagram

connection are entirely based on the frequency of communication and interaction, which conforms to our previous analysis. Active contacts are also useful in distinguishing real users from fakesters. Real users do not normally communicate with fakesters. Thus, fakesters are often in the state of being inactive. If real users do communicate with fakesters, then fakesters turn to be active. This is the case where authentic users employ fakesters as their online personas.

5.3.4   Preferential Attachment: Beyond Reciprocity

People with higher social capital can attract more people with less effort than those with lower social capital. This is called *preferential attachment* and is considered as one of the two important factors in the Barabasi-Albert model which theorises that complex network exhibits power-law degree distribution. Because the social capital accruing from a relationship is much greater to the extend that the person who is the object of it is richly endowed with capital, the possessors of an inherited social capital, symbolised by a great name, are able to transform all circumstantial relationships into

lasting connections[15]. They are so well known that they do not even have to make the acquaintance of the acquaintances.

The power-law degree distribution of social network, which is due to the effect of *preferential attachment*, reflects our living experience that the rich get richer. In the study of distribution of wealth, 20% of the population in a given society hold 80% of total wealth of the society. The top rich people can grow even richer by taking advantage of their existing capitals. The topology can also explain the spreading of disease. Research has found that there is no epidemic threshold for virus to spread all over the network with power law degree distribution, regardless the rate of infection[102]. This implies that power law structure can facilitate the dissemination of information and knowledge. As long as these information exist, they will eventually spread all over the network if they have any value.

We therefore propose that individuals with more connections should be able to acquire new connections with less effort. There should be no universal frequency of communication but a decreasing range of frequency over the individuals with increasing connections. The use of universal frequency was a sound decision at the first thought. It is reasonable to argue that the more one communicates, the more connections that one can establish. Some take for granted that there is a linear relationship between the number of contacts and the effort one has spent on social grooming. However, this is not true. If we look at the topology of the social network into account, the hallmarks of any social network of human beings are power-law degree distribution and assortativity. The power-law degree distribution, or scale-free character, is the base for the social network to spread the information and knowledge quickly. It also reflect the self-organised feature of the network that is robust and resistant to random attack. Unfortunately, the message network of online social network does not follow the power-law degree distribution. As a case in point, the network of testimonials on Cyworld exhibits exponential degree distribution, rather than the power-law degree distribution as presented in the real-world social network, as indicated in Figure 5.2.

Research on Facebook indicates that the probability distribution of number of messages sent per user does not show power law distribution either, as illustrated in Figure 5.3. It is in-between the heavy-tail Pareto or power-law and thin-tailed exponential distribution in terms of its asymptotic behaviour[53]. This is also true for the social network of the University of Southampton. There are about 10% of people have number of connections above 500 and 30% of them have number of connections between 200 and 500. However, it is untrue that all of them are actively engaging on Facebook online activities such as public wall posting, photo and video commenting, gift exchanging and blog commenting. On the contrary, most of these users are fairly inactive and rarely engage in these social grooming activities. But there are indeed a significant portion of people will actively engage with these activities. For these people, the group with connections above 500 will have exchanged information slightly more than those with connections between 200
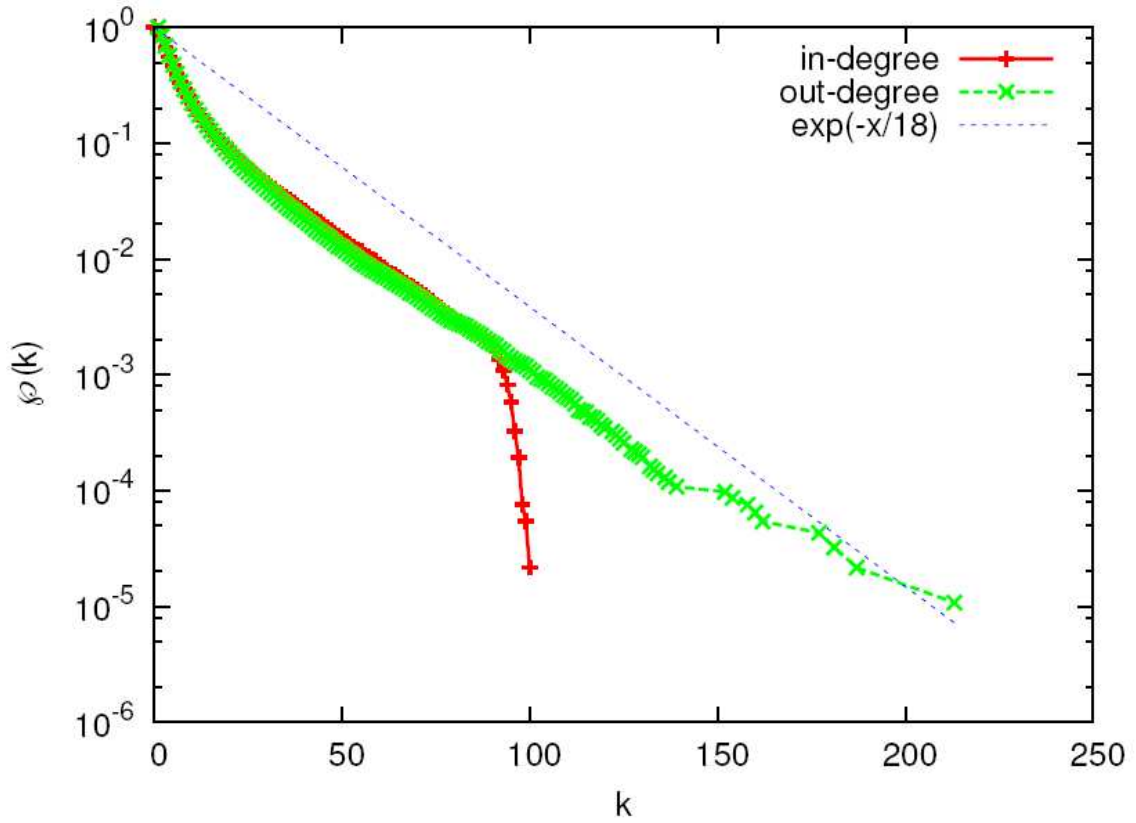
FIGURE 5.2: Cumulative distribution of in-degree and out-degrees of Cyworld's testimonial network. Picture taken from [4]

and 500. It is unlikely for people with more connections to spend equally exponential time and effort for social grooming. It is not achievable even if they spend all of their total time.

The statistics confirm that the social network purely based on universal frequency of communication will not exhibit power-law degree distribution. That is, f_k is not a constant. In fact, as we argued earlier, highly connected users simply do not have sufficient time for such expensive social grooming. In order to design an algorithm to transform the message network to a network with complex network topology, we would like to find out the mapping between the degree and its frequency of communication. First, the power law distribution goes as follows:

$$p_k \sim k^{-\gamma} \qquad 2 < \gamma < 3 \qquad (5.1)$$

Second, we need to find out the relationship between the p_n and number of messages. In our study on the social network of University of Southampton, we are not able to access either the data of private messaging or those of instant messaging. Therefore, we do not have first-hand data to draw the connection. However, according to the research
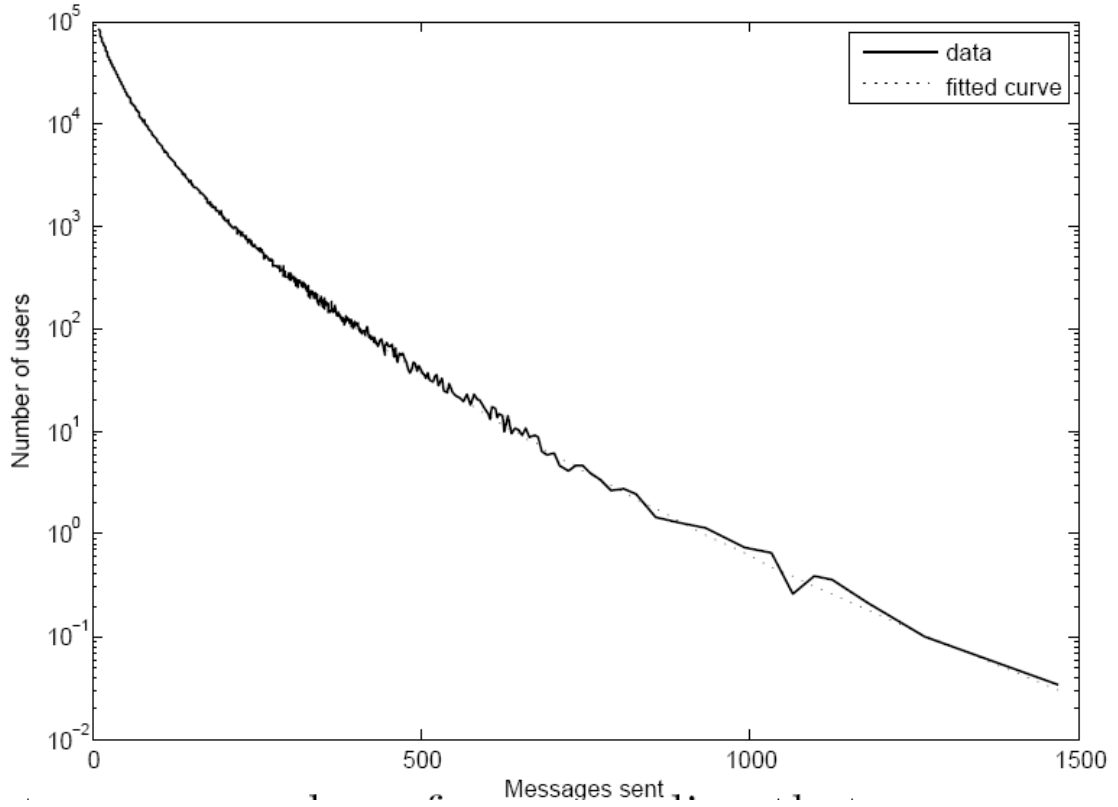
FIGURE 5.3: Number of messages sent versus number of users sending. Picture taken from [53]

of Golder et al., who did manage to access the Facebook messages, the relationships between p_n and number of messages goes as follows[53]:

$$p_n \sim n^{-\alpha n^{\beta}} \qquad \alpha > 0, 0 < \beta < 1 \tag{5.2}$$

Given the formula above, we introduce the frequency function f_k, so that the previous distribution becomes a power law distribution. Replacing n by kf_k in eq.5.2 and making index equivalent to γ, we have:

$$\gamma \sim kf_k{}^{\beta} \tag{5.3}$$

Solving the eq.5.3, we have:

$$f_k \sim \frac{\left(\frac{\gamma}{\alpha}\right)^{\frac{1}{\beta}}}{k} \tag{5.4}$$

Thus, the relationship between frequency and degree is:

$$f_{k+1} \sim \frac{k f_k}{k+1} \tag{5.5}$$

Eq.5.4 suggests that the frequency of communication is inversely proportional to the degree. In another word, the more friends we have, the less effort we need to spend on communication and social grooming. This appears to be true by our intuition. People who are highly popular and opinion leaders can maintain tens of hundreds of times of meaningful connections that ordinary people can manage to establish. However, while they are very likely to spend more time on social grooming than most of ordinary people, they are very unlikely to spend tens of hundreds of times of the time and effort of ordinary people for social grooming. If we put the value for ordinary users on social grooming to be 20% of their total time, then people who are highly connected and opinion leaders can only spend five times of that in maximum. If ordinary users can have the number of connections at 150, as described by the rule of 150 or Dunbar's number, then people who are centres and hubs of the social network can only have 750 in maximum. This figure is much smaller than what we have found on most social network sites, where users with more than one thousand connections are not unusual. Although, as we analysed earlier, a significant part of these users are friendship collectors, fakesters and fraudsters and even spammers and phishers, there are indeed a decent part of them are genuine popular figures who enjoy a high reputation in the social network. These are usually bloggers, journalists and celebrities. They may simply acquire connections passively without much effort. The inverse relationships between frequency of communication and the degree of individual users, as we concluded earlier, explain the fact that these *Facebook whales* demand less time than ordinary users to acquire more connections, as they can leverage the connections they have already acquired.

Because the method reflects the real activities of genuine social network, the configuration can single out the highly connected nodes from the whole pool of nodes. With the information hubs and opinion leaders that have more connections, information and knowledge can spread much faster and further.

5.3.5 Assortativity

A unique feature to social network is that the degree correlation is positive, which is called assortative mixing or assortativity. The characteristic is first proposed by Mark Newman and is then confirmed by numerous research on the topology of real-world social network. This is in contrast with other non-social networks such as information network, technology network and biological network. In non-social network, new nodes will simply follow the rule of *preferential attachment* and connect to old nodes with bigger number of degrees. However, in social network, new nodes will not only follow

the rule of *preferential attachment*, but also exhibit the pattern of *assortativity*. This means nodes with similar number of degrees will connect to other nodes with similar number of degrees. Hence, in the real-world social network, highly connected nodes will connect to other equally highly connected nodes while less connected nodes will connect to other less connected nodes. The fact that in social network, people connect to others who have similar degrees is echoed by the proverb that birds of a feather flock together, which reflects our living experience. Social network users befriend with whom they share common interests, ideas and values. People with similar social and cultural background can usually foster stronger relationships. For example, in our study of Facebook users of the University of Southampton, if we compare the social network of the university users with connections between them and the social network of both university users and those from outside the university, then we can find the assortativity is usually three times larger in the former case. This is largely because university users generally befriend with other university users along the line of similar subjects, schools and departments and other interest groups.

The presence of assortativity in the social network plays an important role in information spreading and knowledge sharing. Information and knowledge can travel much faster in social network than non-social networks such as technological network, information network and biological network. Research has confirmed that assortativity, together with power law degree distribution, can further advance the dissemination of information[96]. This is because highly connected nodes will connect with other nodes which are equally highly connected. They will form a core group of highly connected nodes, which could serve a "reservoir" for information, sustaining an epidemic spread. A message originated from one of these nodes can spread across the whole "reservoir" in a short time. It will then travel from highly connected nodes to other ordinary nodes. The process is equally true if the message is originated from the less connected nodes according to *preferential attachment*, this less connected node will normally connect to a node which is highly connected in the social network. When this less connected node broadcasts a message in his or her ego-centric network. The message will be received by highly connected node. This node, which is connected to other highly connected node in the "reservoir", will pass the message to other members of his or her ego-centric network, where a significant number of its members are highly connected nodes. The message can then propagate along the chain of highly connected hubs and eventually reach every node of the social network. This is different from other non-social network. For example, on the World Wide Web, many search systems has indexed a huge number of websites. Information will travel from these lowly connected site to the search engine, which is highly connected. However, because this search engine is not connected to other search engine or web portals which are equally highly connected. The information will only stay locally within that particularly Web search engine and its ego-centric network.

To take advantage of assortative mixing in information spreading and sharing, we decide

to incorporate this feature to *ActiveLink*. There will be two tiers of frequencies of communications. The upper tier frequency, which is smaller than the lower tier frequency, is only applied to users with degrees above the threshold level, which is always less than the median of the total degrees. The mechanism of two tiers of frequencies is actually a logical step to *preferential attachment* because if nodes in the upper tier attempt to befriend with those in the lower tier, the connections will be directional. However, in the social network, the connections are always mutually acknowledged and therefore bidirectional. The values of first and second tier frequency are approximated by trial and error method. We can first assign an initial value to the first tier frequency and then assign another value which is smaller than the first one to the upper tier frequency. The topology of the resulting social network based on the two-tier active connections will be measured against that of real-world social network. If it follows power-law distribution, then we use the second frequency values. Otherwise, we adjust the second tier frequency by adding or reducing one, then we will measure again the topology of the resulting social network, until it conforms to the topological feature of real-world social network.

5.3.6 Representative Democracy Model

The mechanism of forming active connections can be understood by using representative democracy model. Representatives are elected from their respective constituencies. In each constituency, there are several candidates competing for the election. Candidates spend a lot of time and efforts to talk to the voters for political campaign. Once they are elected and become members of the parliament, they can acquire more connections with different representatives and celebrities in the society. However, they will still need to maintain close relationship with their constituencies. Otherwise, they will be distant from their voters and may lose their votes in the next election. If they lose their seats in the parliament, they may lose those connections with other representative and celebrities. Here, there are two tiers of network: the voters and the MPs. The first tier frequency goes between voters and voters and voters and MPs. The second tier frequency goes only between MPs and MPs. Both groups need to spend time and efforts on social grooming in order to keep the connections alive. However, MPs have an advantage over voters in that they can leverage their social capital to achieve more social capital. Ordinary voters who want to be MPs to leverage the social capital must work hard to reach the threshold amount of number of connections.

The model is also consistent with the idea underpinning Watts-Strogatz model proposed by Watts and Strogatz[123]. WS model has its roots in social systems in which most people are friends with their immediate neighbours on the same street, colleagues, people whom their friends introduce them to. However, everyone has one or two friends
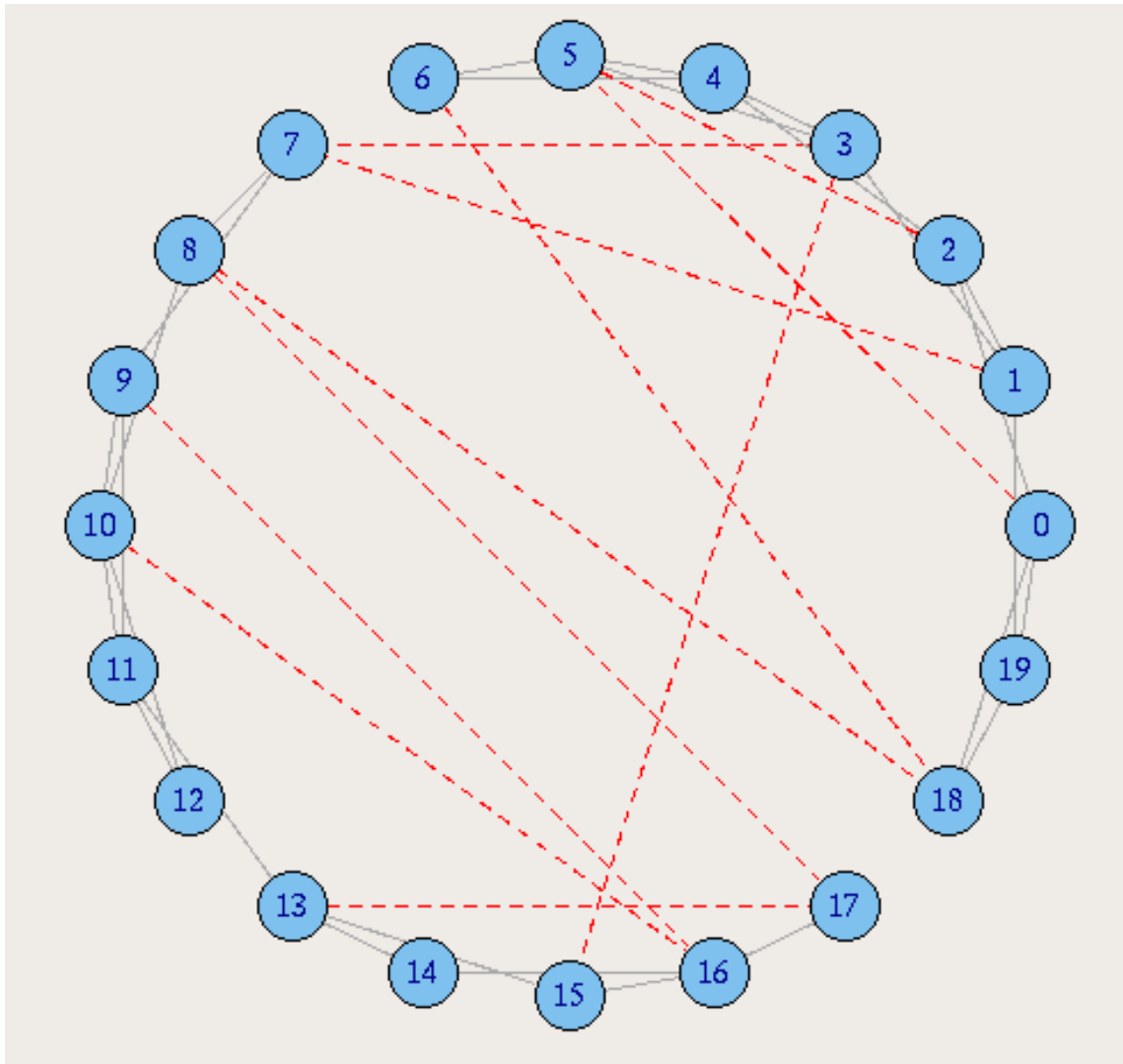
FIGURE 5.4: Illustration of social graph that is identified by ActiveLink Algorithm. Hard lines represents immediate neighbours. Dashed lines represent long-range contacts.

who are a long way away – people in other countries, old acquaintances – who are represented by the long-range edges obtained by rewiring in the WS model[5], as shown in Figure 5.4. People generally spend more time with their immediate neighbours and less with long-range contacts. Here, the first tier frequency goes to people who have only local connections. Second tier frequency goes only to people with remote connections. Both groups need to spend time and efforts on social grooming in order to keep the connections alive. However, people with remote connections have an advantage over people with only local connections in that they can leverage their social capital to achieve more social capital. People with only local connections who want to establish remote connections to leverage the social capital must work hard to reach the threshold amount of number of connections.

5.3.7 An Algorithm for ActiveLink

Given the previous analysis, we present an algorithm for the calculation of active links. Let D be the period for connection expiration, S be the set of all members of the social network and f be the frequency of information exchanging such as private messaging and mutual public wall posting, m be the median of connections, S_1 be the set of less connected nodes, S_2 be the set of highly connected nodes, C be the connection cap, m_0 be the median of connections of the previous iteration we have:

Algorithm 2 ActiveLink: Identify Meaningful Social Connections

---

Input: A Social Network S
Output: A Social Network S' based on Reciprocal Communication
ACTIVELINK(S, D, C, m_0)

1: **for** Every D days **do**
2:      $S' \leftarrow \{\emptyset\}$
3:      $f = 2$
4:      **while** $S \neq \{\emptyset\}$ **do**
5:          **loop**
6:              Apply f to $S \Rightarrow m, S_1, S_2$
7:              **if** $m > C$ **then**
8:                  $f = f + 1$
9:              **else if** $m < m_0$ **then**
10:                  $f = f - 1$
11:              **else**
12:                  *break*
13:              **end if**
14:          **end loop**
15:          Calculate Mean Degree k_1 of S_1 and k_2 of S_2
16:          $f = f k_1 / k_2$
17:          $m_0 = m$
18:          $S \leftarrow S_2$
19:          $S' \leftarrow S' \cup S_1$
20:      **end while**
21: **end for**

When applying f to S, we get temporary active connections for each member. The median of the distribution of these connections is denoted by m and those who have less than m active connections belong to the first tier network S_1 while others belong to the second tier network S_2. m_0 refers to the median of previous calculation. The algorithm adjusts the value of m so that $m_0 \leq m \leq C$. Then, the second tier frequency of communication is estimated as $f = f k_1 / k_2$, which is applied to the second tier set of members.

Depending on the requirement and accuracy, the network may be divided to quartiles and even more sections. In our implementation, we will employ a topology calculation module to check if the resulting degree distribution fits into our expectation.

5.4 Summary

The chapter introduces the algorithm of *ActiveLink*, a sophisticated reciprocal method aiming to identify meaningful connections in the social network. We gave a detailed description of what affects active links and how it works. The algorithm goes beyond simple algorithms based on reciprocity. In stead, it takes advantage of the social capital that a user has acquire when identifying meaningful connections. We argued that the algorithm is capable of recognising long-range contacts who often communicate less frequently than immediate neighbours such as friends, colleagues and schoolmates.

Chapter 6

Experimentation and Evaluation

6.1 Introduction

In the previous chapter we presented the algorithm of *ActiveLink*, which aims to identify meaningful online social connections. Meaningful connections refer to friends on SNSs whom users have actually met offline or those with whom they maintain regular contact. The social graph identified by the algorithm is the foundation of a social network site as it represents the real network instead of the inflated one. In order to increase our confidence in the capability of the algorithm and to support the theory of friendship inflation, we need to validate the ActiveLink algorithm experimentally.

In order to perform such an empirical validation on a reasonably large scale based on real-world example, we will use the social data of Facebook users of the University of Southampton. The university has about 24,735 students and around 5,000 staff[1]. Only users with a university email account can join the network of Uni. Southampton on Facebook. As of this writing, there are 24,512 people in the Uni. Southampton network, according to Facebook's statistics. We managed to harvest 22,553 users with their profiles and connections between them.

In this chapter, we describe the how we apply the algorithms to the data and present the results of the empirical study. Three reference algorithms will also be presented to compare against our ActiveLink algorithm.

---

[1] http://www.hesa.ac.uk/dox/dataTables/studentsAndQualifiers/download/institution0607.xls

6.2 Approach

Our methodology imitates the behaviours and activities of a user with an email account of the University of Southampton. He or she can browse other users' profiles and connections within the same university network, subject to individual user's privacy settings and other preference configurations. The approach is different from general Web scraping method in that we need to login on the website (Facebook in this case) in order to harvest the website. Also, because Facebook imposes a stringent regulation on computing scripts, we need to make our harvesting script behave more like a human being. Our general approach to verify the ActiveLink algorithm is as follows.

We begin by harvesting the social graph of the University of Southampton. Noted that the connections of these networks only consist of those among university users. This means the connections between users from outside the university are not counted. The resulting social graph is essentially a large ego-centric network of the University of Southampton. Personal details about users will be used and analysed. Communications include the mutual exchange of virtual gifts, public wall posts, comments on notes (Facebook-style blogging), photos, videos and posted items are also relevant. Secondly, we will prepare the experiment data by removing the multiple and redundant personal information and complete the connections which may be directional rather than mutual. Thirdly, we run several simple reciprocity algorithms to identify social networks as our reference graphs. The first reference algorithm considers one-way communication between any users. The second reference algorithm is based on reciprocal communication where users in both parties have exchanged information at least once. The third reference algorithm requires users exchange information at least twice. Finally,we will apply the ActiveLink algorithm to identify meaningful connections. We will compare the network that is identified and extracted from our algorithm and from a reference algorithm, as well as the original inflated network. The topological properties to be compared include degree distribution, average path length, clustering coefficients, assortativity, etc.

In this empirical study, we are attempting to (1) confirm the theory of friendship inflation by contrasting the original social network with the graph generated from communication networks; (2) verify that the ActiveLink algorithm can identify long-range contacts which can not be captured by simple reciprocity algorithms. It is our intention to study the scalability of this algorithm, but in the case of acquiring large amount of data this was not possible within the limits imposed by the regulations from most social network sites and by the restrictions from users' privacy settings and preferences. For example, in our experiment, we acquired our data from the network of the University of Southampton on Facebook. This requires a valid University email address, which usually ends with "soton.ac.uk". In addition, Facebook imposes a strict limit traffics on the site and therefore the harvesting script has to run slow enough so that it will not be detected by Facebook's monitoring program. The experiment can also be deployed on other networks

such as the region networks of Portsmouth and London, which have far bigger population than that of the University. However, the problem with networks of this type is they are difficult to be verified in the offline world. Also, the interaction between members in these networks appears less intensive than those taking place in the University.

6.3 Data Acquisition

We start by harvesting the social graph of the University of Southampton on Facebook. The harvesting algorithm starts from an arbitrary node and runs a breadth-first search in the network. The algorithm will include all the nodes from the university, however, it will exclude those from outside the university. There are several categorisations of contacts such as friends from schools, companies, different geographical networks and those updated recently. Our algorithm will take the connections but ignore these categorisations. In the University network, the default privacy settings are to allow members of the same network (in this case, the University of Southampton) to view others' profiles. As the network grows, more University users begin to recognise the privacy issues in the social network and change their privacy settings. However, in most cases, members of the same network can still send a private message to each other and view each other's friend lists. The service to navigate through other's social network, even though they do not have direct friend connections with the view, is important to us as it is possible to use a snowball sampling method to crawl the whole social graph of the University network.

This snowball sampling algorithm is arguably the only feasible method to crawl all the data in the network of the University for the following reasons. First, we do not have a list of UIDs of all the University users and therefore we can not index the friends of users by leveraging Facebook APIs. Second, the method makes sure the resulting social graph is a connected component. Third, Facebook provides APIs for accessing users' data, however, without knowing users' Facebook UIDs it is impossible to verify if any two people are friends. Even we have these UIDs, the verification function is painfully slow. It takes much longer time to crawl the social graph. The algorithm is given in Chapter 4. In our experience with Facebook, this adapted breadth-first search method will pick up highly connected nodes or hubs very quickly in the beginning.

While the Facebook APIs are inefficient in acquiring the social graph, they are however suitable in retrieving profiles of the university users. These profile items, subject to users' privacy settings, include user's self description, activity descriptions, affiliations, birthday, books, current location, education history, email addresses, hometown location, interests, quotes, status, timezone, favourite TV and films, work history, etc. The information are set to be visible to members in the same network of Southampton University by default but it is not uncommon that users change the privacy settings to

restrict the visibility of some personal information. Facebook has provided both APIs and SQL query language for accessing users' profiles[2].

Our algorithm can acquire the information that is visible to other University users. Given the UIDs that are harvested from the snowball sampling method that is described above, our algorithm can retrieve the Facebook profiles by only using the services provided by the Facebook platform. Figure 6.1 shows some demographic information about the University users such as age and gender. There is about 53% of male users and 47% female users. The vast majority of users age between 18 and 24. This suggests that most Facebook users in the University network are undergraduates. For the age below 18 and beyond 30, the population quickly shrinks to a very small number, which means there are few University staff using the Facebook at the point of our experiment. However, we believe that as the network grows, there should be increasingly more University staff using Facebook as of this writing.
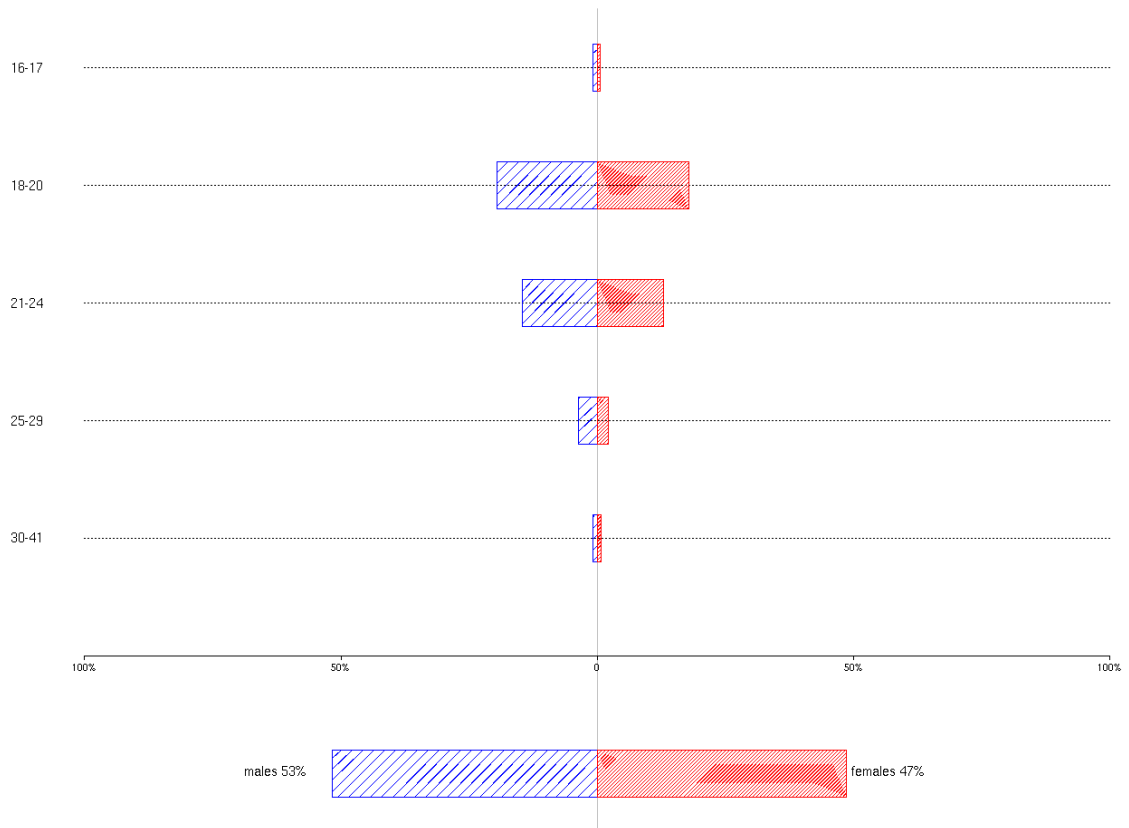
FIGURE 6.1: Sample Demographics: Age and Gender Distribution of the University users.

Another set of information that is relevant to our experiment is users' activities, in particular, the interactions between users. While we can not access the direct communication between Facebook users via private messaging and instant chatting, we can,

---

[2]http://wiki.developers.facebook.com/index.php/Main_Pag

however, access a broad range of direct and indirect communications that meant to be visible to other members in the university network. These communications include the mutual exchange of public virtual gifts, public wall posts, comments on notes (blogs), photos, videos and posted items. There is also another type of communication which requires the installation of third party applications, such as superpoke. These type of communications may not be universally available to all the users and therefore are not used in our experiment.

The following algorithm details how we harvest the information about photo comments, one of the six sources of users' interactive activities. First, we retrieve the 22,553 UIDs from the database and put it in an array. These UIDs are crawled from the snowball sampling method and are the Facebook identifiers that represent the users in the University. More than 10,000 UIDs beginning with "28610" or "50", which indicate the early adopters of Facebook in the University. Then, we take a UID and retrieve the first page of photo commenting for that user. The page will display how many posted photos in total. There are two cases: (1) if the number of posted photos is less than 20, then there is only one page; (2) otherwise, there is more than one page. In the case of only one page, we will further look at how many photo comments have been made for each posted photo. UIDs of users who have made these comments will be harvested from the profile links. In the case of more than one page, after we finish the iteration as described in the former case, we will advance to the next page by accessing the next 20 posted photos. We then perform the same harvesting process in this new page. This is iterated until no more posted photo is found

---

Algorithm 3 Harvest Photo Comments From Facebook

Input: An Array UIDs of University Users on Facebook
Output: Photo Comments on Facebook
ADAPTED-WEBSCRAPING($array(uid)$**)**

  1: Login on Facebook
  2: **for** Each Uid of Array(UID) **do**
  3:      Fetch the first page of photo commenting
  4:      **if** There is only one page for photo comments **then**
  5:          Harvest UIDs of friends making comments
  6:      **else if** There is more than one page **then**
  7:          Harvest UIDs of friends making comments by page
  8:      **else**
  9:          *break*
 10:      **end if**
 11:      Sleep(10)
 12: **end for**

These data will be checked for integrity before they can be used in the experiment and data analysis. One particular important area to look at is the directional connections. Some users who have adopted more rigorous privacy settings may not be reached directly, yet they may be reached by some of their friends who have less restricted settings.

Therefore, the raw social graph as harvested from Facebook will have some directional links, rather than mutual links. In this case, we will complete the connections by adding the complementary directional links. The original data of users' activities include a lot of information published by users from outside the University. In fact, there are more data posted by users from outside the University than the University users. Thus we need to distinguish these two set of data and only select those published by the University users. For the data of interactive activities, we also need to exclude two set of noises: comments made by the same user who posted the information and comments made by users from non-University users.

6.4   Data Analysis
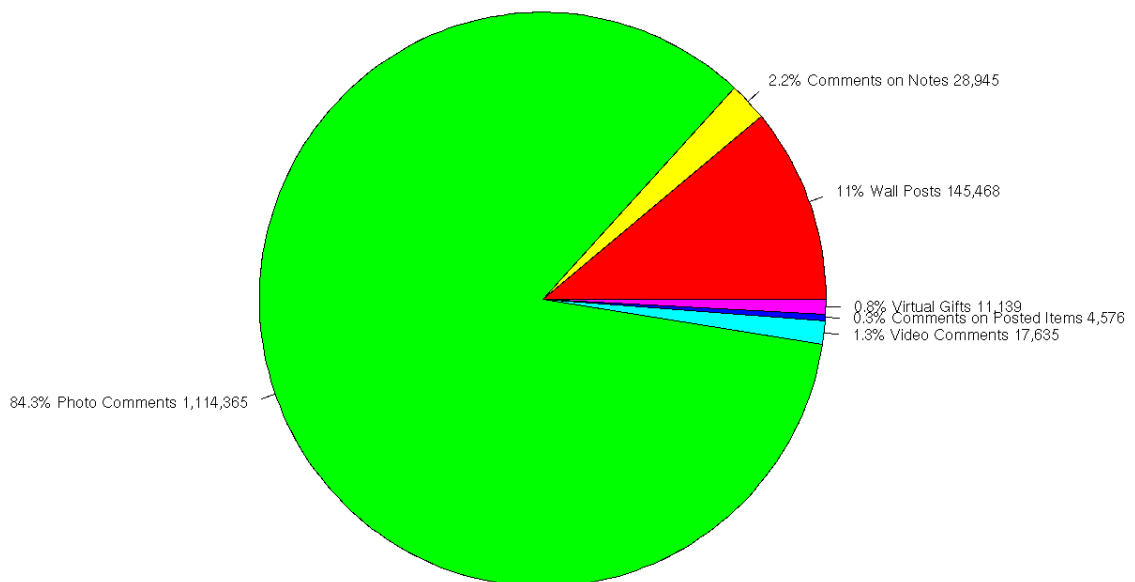
Figure 6.2: Six Sources of Interaction Activities.

Figure 6.2 indicates the six sources of interaction activities. It shows that the dominant activity that we are able to harvest is photo comments. With more than one million photo comments, it represent 84% of total activities. Wall posts, which takes about 11%, come at the second place. The surprising discovery is the number of comments that have been made on the notes, which are Facebook-style blogs. The number is slightly less than 30,000, responsible for only 2.2% of all activities. Our experience with Facebook suggests that there should be more wall posts than photo commenting. The fact that we harvest more photo comments than wall posts is due to ajax technique that Facebook has used for viewing wall posts. Our web scraping algorithm can only harvest the first page of wall posts whereas it can harvest the photo comments page by page. As a result, it sees more photo comments than wall posts. This does not mean

the activity data we acquire can not represent users' overall activities. Users involve with photo commenting will general publish more wall posts and comments. Hence, we consider the photo comments of more than 1 million, together with other sources of activity information, is appropriate for our experimentation and evaluation.
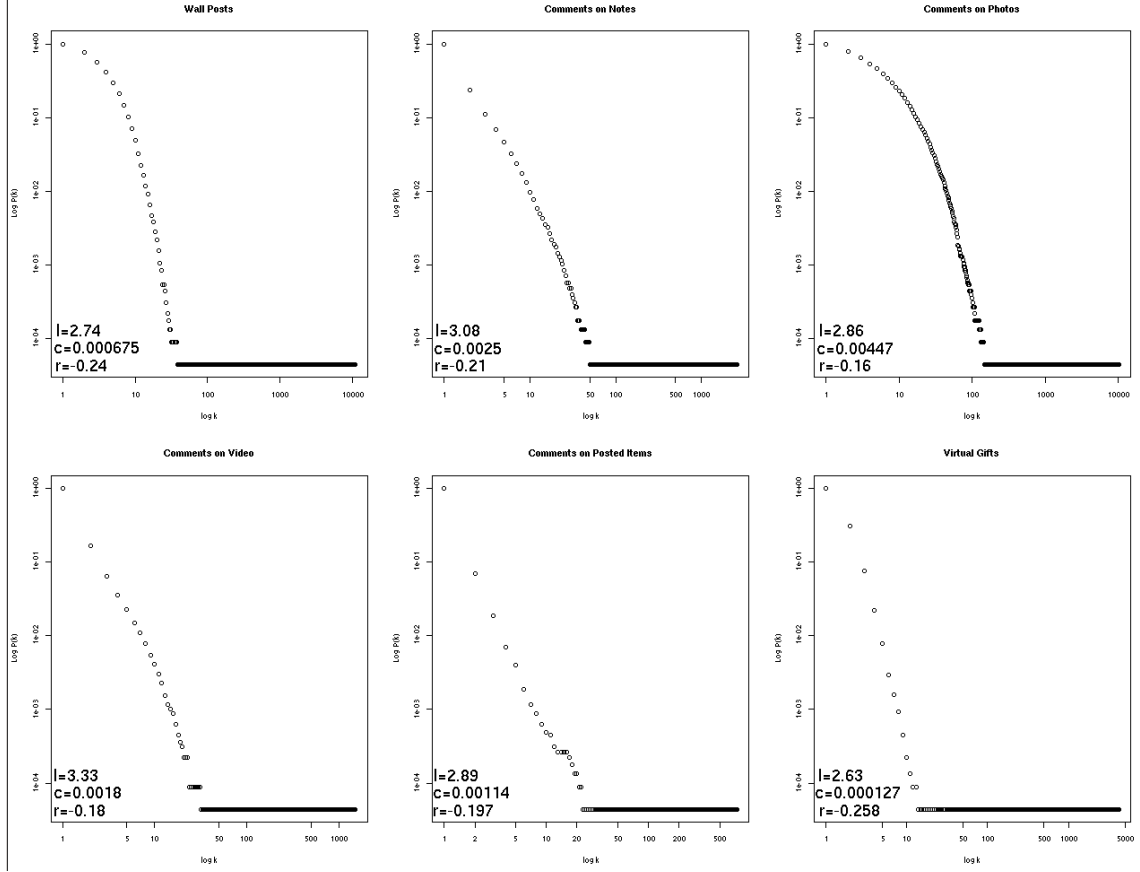
FIGURE 6.3: Topological Characteristics of the Communication Networks.

Figure 6.3 shows the topological characteristics of the six sources of interactive activities. The topological properties we will investigate include degree distribution, average short paths, clustering coefficients and assortativity. The degree distribution of these graphs roughly takes the shape of power-law degree distribution. It is notable that the degree distributions of wall posts and photo comments exhibit cutoffs that resemble real-world social networks where there are age or capacity constraint. The average path lengths are around 3, suggesting that these communication networks are quite small. However, the clustering coefficient is near neutral and the assortativity of them is unanimously negative. The zero clustering coefficients and negative degree correlation pattern suggests that online activities concentrate on a small amount of active users.

6.5   Experiments

After an initial data analysis, we will carry out some experiments in this section. The first experiment is to extract the one-way communication network from the online social network of the University. We compare the original online social network and the one-way communication network. The second experiment is to apply simple reciprocity algorithms to the communication network, resulting in different social graphs with different frequencies. These social graphs can serve as benchmarks for the test on our algorithm, which will be carried out in our last experiment. We are particularly interested in how the ActiveLink algorithm is capable in identifying long-range contacts that are not able to be captured by other reference algorithms.

6.5.1   One-Way Communication Network

Topological Characteristics of One-way Communicatiom Network
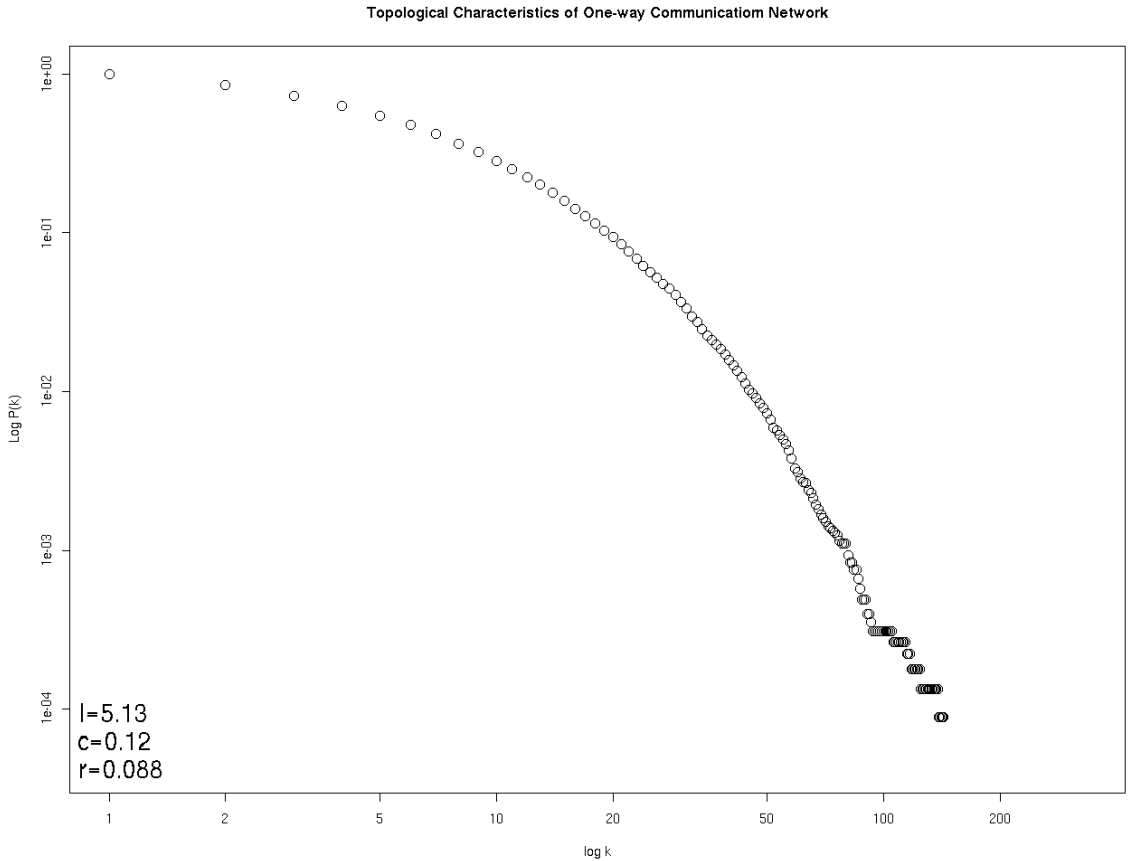
FIGURE 6.4: Reference Algorithm: One-way Communication Network

The first experiment is to identify the one-way communication network based on the six sources of interactive activities. The experiment treats all the sources of data equally and does not assign any weight to different activities. Connections will be placed between any two nodes if they have communicate with each other at least once and are friends

with each other. The resulting network, which is essentially the union of the individual directional communicative network of the interactive activities, remains directional. The topological properties of the network can be found in Figure 6.4.

The range of the x-axis of degree distribution is between 1 and 160, i.e., $1 \leq k \leq 160$, which has shrunk dramatically from the range of degree distribution of the inflated online social network, $1 \leq k \leq 900$, as in Figure 4.16. The population of one-way communication network is 11,980 as shown in Table 6.8, much smaller than the original size of the social network, which is 22,553. This is consistent with the statistics from Facebook Data Team[3] who found that the size of communication network is much less than that of the original social network. Communication only takes place between a small fraction of the whole friends' social network. A significant part of online connections are the result of people being silently linked to others. The one-way communication network is closer to the real-world social network, as there is a clear cutoff of the degree distribution. It has, however, a small assortativity and clustering coefficients but larger average path lengths. It shows that without the inflated number of friend connections, the one-way communication network appears to be sparser. The significant statistical difference between the original online social graph and the one generated from this algorithm, together with the theory of Dunbar's number, supports the previous analysis that there is a huge friendship inflation in online social networks.

6.5.2   Simple Reciprocal Network

The second reference algorithm is a simple reciprocity algorithm where the connection is identified if (1) people involved in this connection have already established an online friend connection by using *static link*; (2) they have exchanged messages at least once. For our convenience we choose the two frequencies, f=1 and f=2. We think this is sufficient for us to carry out the benchmark algorithms as when f>2, the network is too small to see any significant effect. The resulting social graph is a simple reciprocal network, where the topological characteristics are illustrated in Figure 6.5 and Figure 6.6, respectively.

The ranges of the x-axis of degree distributions of f1 network is between 1 and 40, i.e., $1 \leq k \leq 40$, a further contraction from both the original social graph and one-way communication network. The range of the x-axis of degree distribution of f2 network is between 1 and 30, which is even smaller than the f1 network. The population of the two network even reduces to 7,397 and 4,389 respectively, a substantial cut from the one-way communication network. This suggests that reciprocal communication only takes place in a small amount of people who do communicate with each other. The reciprocal networks appear to be more compact than one-way communication network, as they have smaller average path lengths but larger assortativity. Interestingly, the reciprocal

---

[3]http://www.facebook.com/note.php?note_id=55257228858

networks have smaller clustering coefficient than one-way communication network. All these three networks, however, maintains power-law degree distribution. For f1 and f2 network, the topological properties are remarkably similar to each other, except the range of x-axis.
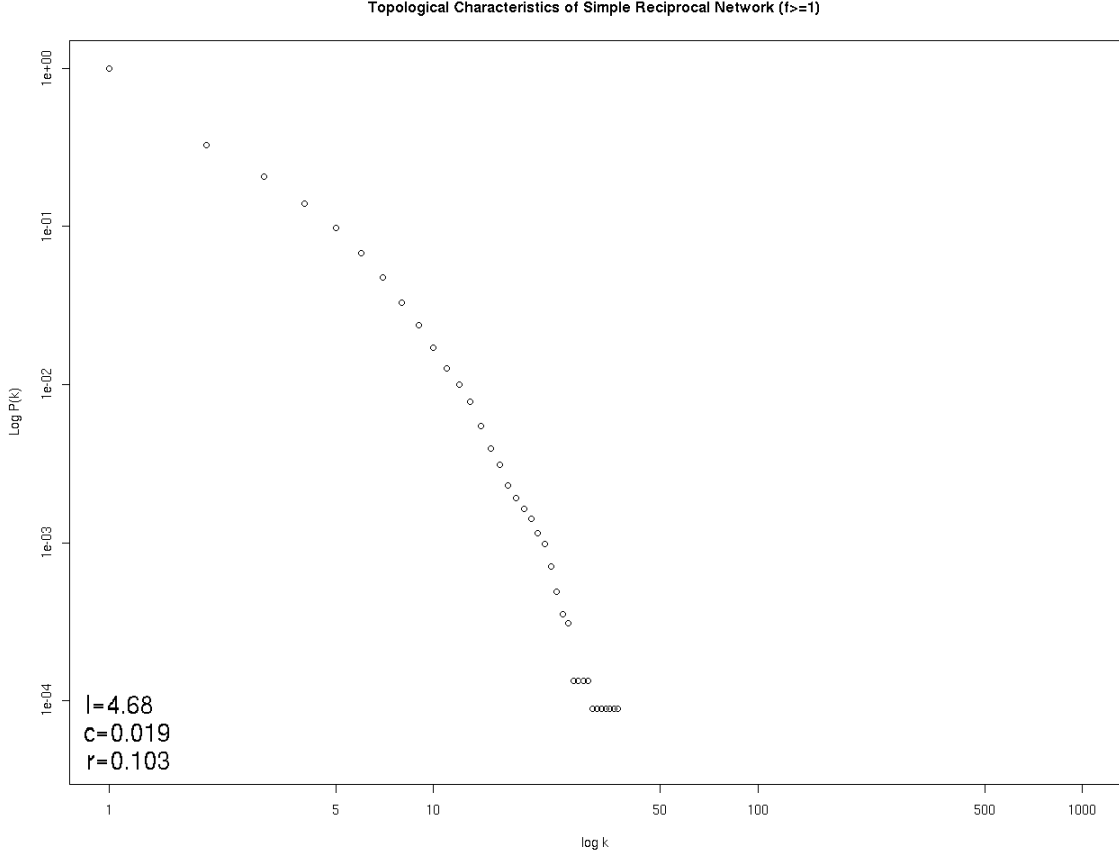
FIGURE 6.5: Reference Algorithm: Simple Reciprocity Algorithm where f=1

6.5.3 Apply ActiveLink to Online Social Network

In this section we carry out an adapted ActiveLink algorithm on the Facebook dataset. We start the algorithm by setting the period for connection expiration $D = 30$, the connection cap $C = 150$ and the median of connections of the previous iteration $m_0 = 0$. With f=2 and then f=1, the algorithm iterates twice for the social network, resulting in the new social graph S'. This adapted algorithm is simpler than the one we proposed in Chapter 5, however, it does demonstrate the key ideas of the original algorithms: *preferential attachment* and *assortativity*, both of which are non-existent in simple reciprocity algorithm, not to mention one-way communication algorithm.

Figure 6.7 plots the degree distribution of the resulting network, which may be called active network for our own convenience. Some topological properties are also presented in the bottom left corner in the graph. Table 6.8 summarises the data set of the active
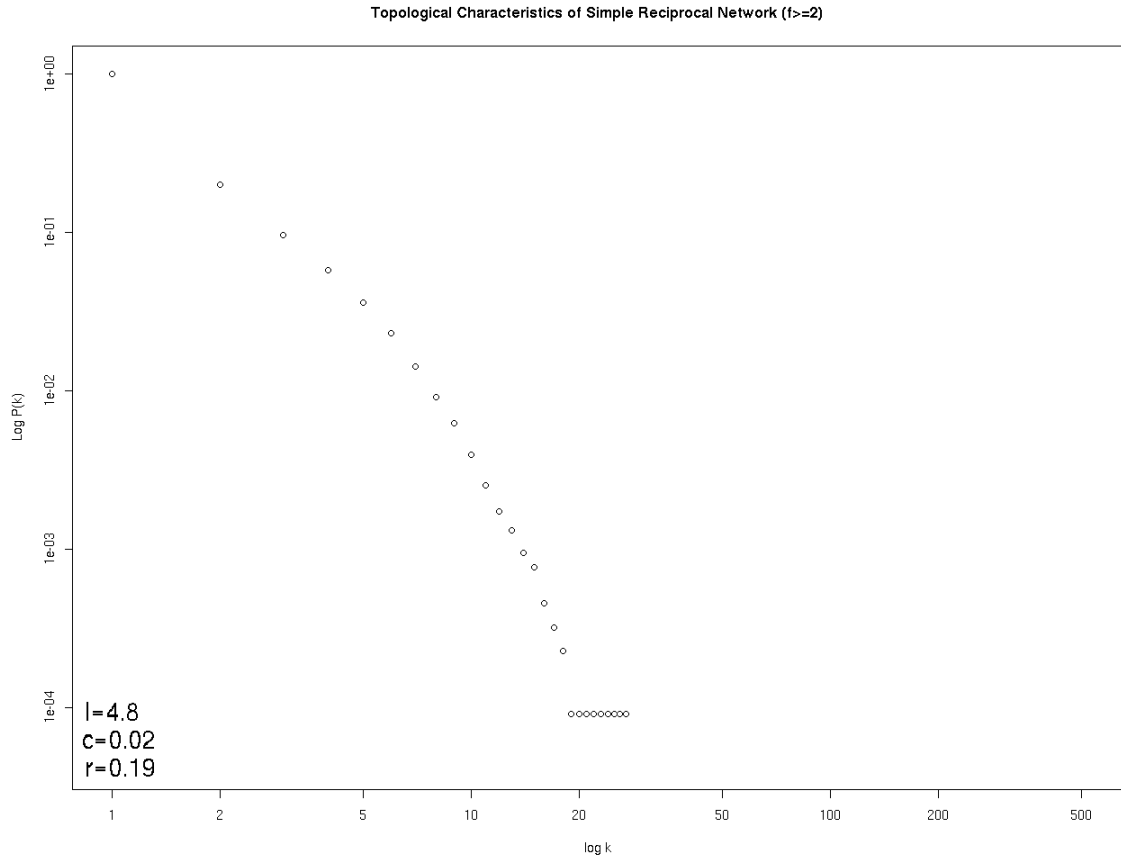
FIGURE 6.6: Reference Algorithm: Simple Reciprocity Algorithm where f=2

network, together with the previous networks. A first look at the table reveals that the population of active network is 6,185, close to the f1 network, which is 7,397, but much bigger than the f2 network, which is 4,399. The average number of friends in active network is 6, which is the same as f1 network, but bigger than the f2 network, which is 4. The range of degree distribution of active network is also similar to that of f1 network, but is bigger than that of f2 network. These figures suggest that whilst ActiveLink algorithm does use higher tier frequency, it can retain connections of lower tier network.

Among other topological properties, the most significant variation in the active network is average path length, which is only 3.75, much smaller than 4.68 in f1 network and 4.8 in f2 network. This suggests there are more shutcuts in active network than other reciprocal networks and therefore information and knowledge can spread faster in it. In the previous Chapter we have stated that this is due to the capability of the ActiveLink algorithm to identify long-range contacts who communicate less frequently than immediate neighbours. To verify the claim, we calculate the average path lengths for various networks and show them in Figure 6.9. In the x-axis, p1, p2. . . represent the length of paths like 1, 2. . . . Bars of different colours represent the numbers of specified path

---

Algorithm 4 Adapted ActiveLink: Identify Meaningful Social Network of Facebook Users in the University of Southampton

---

Input: The Original Social Graph S
Output: A Social Network S' based on Reciprocal Communication
ADAPTED-ACTIVELINK($S, D = 30, C = 150, m_0 = 0$)

```
 1: S' ← {∅}
 2: f = 2
 3: k = 0
 4: while k < 2 do
 5:     loop
 6:         Apply f to S ⇒ m, S₁, S₂
 7:         if m > C then
 8:             f = f + 1
 9:         else if m < m₀ then
10:             f = f − 1
11:         else
12:             break
13:         end if
14:     end loop
15:     f = f/2
16:     S ← S₂
17:     S' ← S' ∪ S₁
18:     k = k + 1
19: end while
20: S' ← S' ∪ S₂
```

---

lengths that are identified by different algorithms, with red for f1 reciprocal algorithm, green for f2 reciprocal algorithm and blue for ActiveLink algorithm.

For path length p=1, the number in active network is smaller than that in f1 network; however, for path length 2≤p≤4, the number in active network is much bigger than that in f1 network. This can be explained in Figure 2.7. F1 network resembles a regular lattice while that of active network resembles a small world network. Thus, f1 network has more connections between immediate neighbours than active network. However, active network has more long-range connections than f1 network due to the rewiring process. These shortcuts can connect nodes from remote distance and therefore shorten the paths between them. As a result, there are more short path lengths in active network than f1 network.

6.5.4 Discussion

We have carried out three experiments in this section, resulting in one-way communication network, reciprocal networks and active network. The topology of one-way communication network confirms our previous analysis of friendship inflation. It also suggest that most people are only silently linked to others but never communicate with
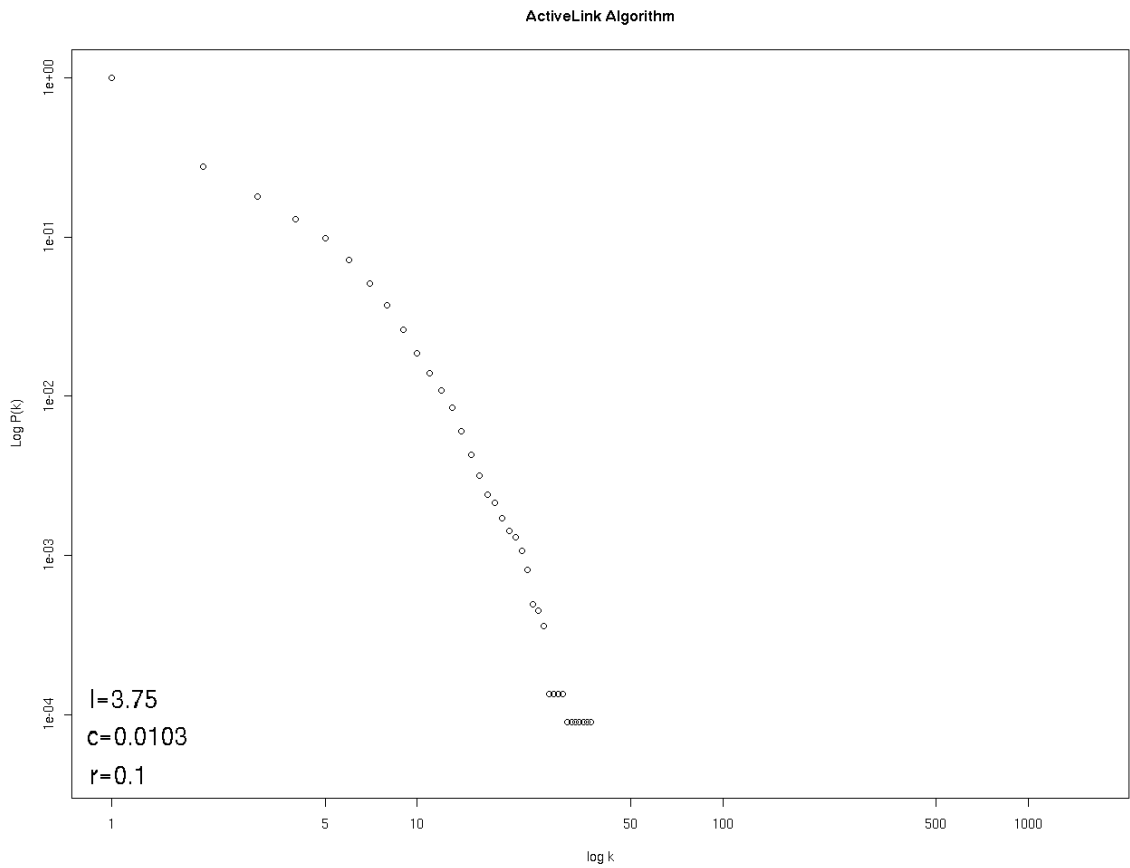
FIGURE 6.7: Social Network Identified by ActiveLink Algorithm

	Population	Avg. Num. Of Friends	Avg. Path Length	Clustering Coefficient	Assortativity
One-Way Communication	11,980	30	5.13	0.12	0.088
Reciprocity Alg. (f>=1)	7,397	6	4.68	0.019	0.103
Reciprocity Alg. (f>=2)	4,388	4	4.8	0.02	0.19
ActiveLink Alg.	6,185	6	3.75	0.0103	0.1

FIGURE 6.8: Summary of data sets from the University networks identified by various algorithms

them. The comparison between reciprocal networks and active networks demonstrates the strength of the ActiveLink algorithm in identifying the long-range connections while retaining other topological properties.

6.6   Summary

In this Chapter, we experiment with Facebook data and gave our evaluated the theory of friendship inflation and the ActiveLink algorithm. It shows that the ActiveLink
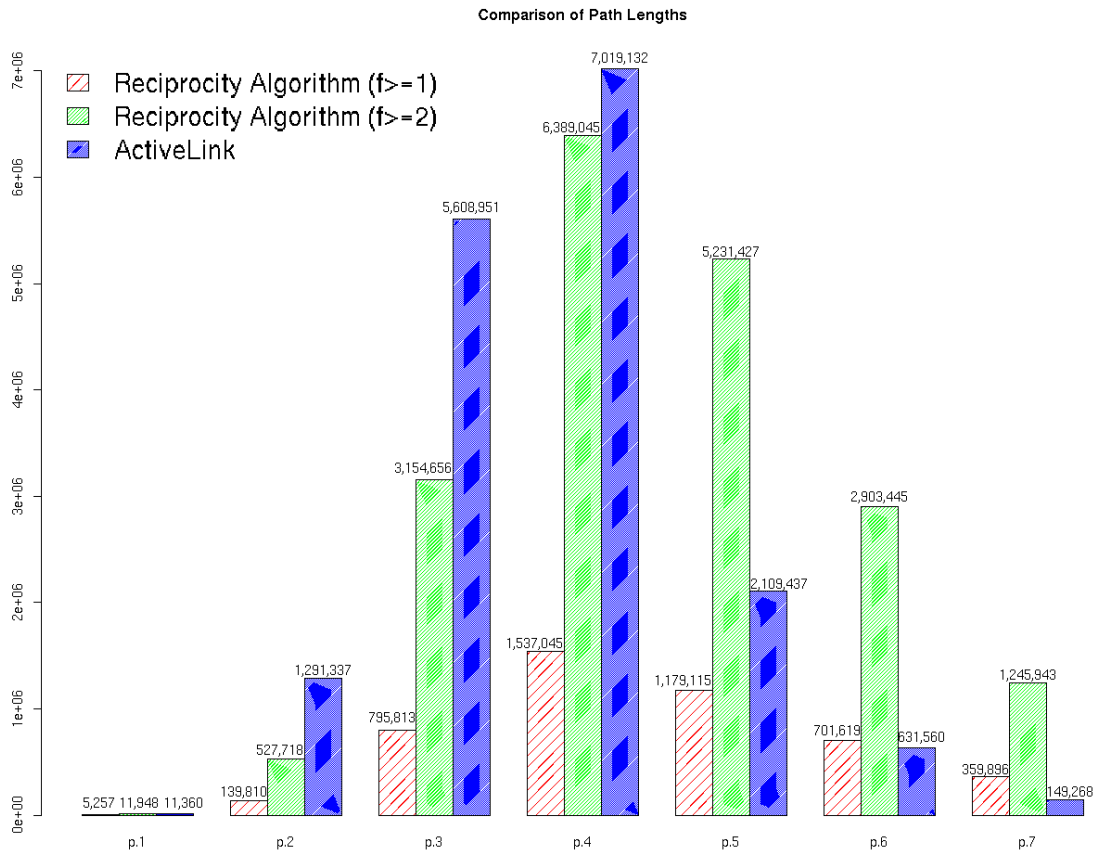
FIGURE 6.9: Shortest Path Length between Each Pair of Vertices.

algorithm can identify meaningful social network by recognising long-range contacts. In the next Chapter, we will propose a RealSpace SNS model based on the active links.

Chapter 7

RealSpace: an SNS Model based on Active Links

7.1 Introduction

In this chapter, we propose *RealSpace*, an SNS model based on evolving active links. RealSpace aims to establish an online social network by capturing the meaningful connections in the real-world social network. The proposed prototype is a social network site with basic functionalities and should be written in PHP. It is essentially based on the LAMP architecture (Linux+Apache+MySQL+PHP). We will first give a high level architecture overview of the system. These include high level abstraction of the system. Then, the structure of component modules is shown to provide some details about the architecture. The system is consist of active links, impression management tools, community identifier, proximity indicator, essential utilities and profile searchings. More details about data schema and applications will also be discussed in the later sections. The application in the outmost layer is meant to include photo and video uploading, chat rooms, news aggregator, online shopping, discussion board, music sharing, etc. There are three other features: *Social Connectivity*, which measures the importance of a user in the social network, *Proximity Index*, which indicates the distance between a user and any other user in the social network, and *Community Structure Detection*, which will identify group structure based on active connections.

7.2 Architecture Overview

Figure 7.1 represents the high level architecture of the RealSpace system. The structure includes three parts: the people's connections that we are going to model; the social network that is based on *active links*; and the applications based on the social network.

In the centre of the graph are the communication and interaction between registered users that the system intends to capture. This may be understood as the "hardware" of the RealSpace social network system because they define the social network, which affects every component of the system. The algorithm for capturing the real-world social network is based on the *active link* as discussed in Chapter 5. The abstraction of social network provides the foundation for the whole system. The second layer is the abstraction of the relationships using evolving social network model. This layer will implement the RealSpace core algorithm which aims to identify the meaningful connections from the original social graph. The techniques that will be used are connection caps, reciprocity, *preferential attachment*, *assortative mixing*, etc. It will include a module that calculates the topological parameters of the resulting social network and compares these parameters with those of the real-world social network. The third layer has essential utilities of the system, such as impression management tools (profile editors), proximity indicator and community identifier. The layer maintains the key applications of the system. Among these applications, proximity indicator and community identifier are based on the algorithm of *proximity index* and *community structure detection*, as proposed in the previous chapter. These graph-based algorithms provide other useful applications and third party applications APIs that can take advantage of the social network. The outmost layer is the various applications such as blogging, video and music sharing. The utilities and applications are modules that can be added to or removed from the system without affecting other modules unless they are interacting with each other. The layer provides the functionalities to the end users and therefore will be crucial to attract users to use the system.

7.2.1 Profile Services

Most social network sites will provide users profile services for them to present themselves. The problem of universal profile and generic persona have been discussed in Chapter 3. The profile services mainly serve as a type of impression management tools. After registration, users will normally be asked to fill in their personal profiles. The information required for the profile include but are not restricted to name, location, hometown, work and education history. The profile can then be viewed by the users, subject to the profile owner's privacy setting. The problem of universal profile is that user can only maintain one profile. Therefore, the content as seen by both their employers and parents will be the same. This will usually cause some social embarrassment and social dramas. To solve the problem, we propose a RealSpace impression management program. The key data structure of the RealSpace social network system is its social network based on *active link*. Every user is treated as a node in the network and is identified by an 8-digit identifier. By using minimum representation of social network users we separate the underlying data structure from the applications. Therefore, we can provide multiple profiles to a user. This is achieved by allowing users with 8-digit
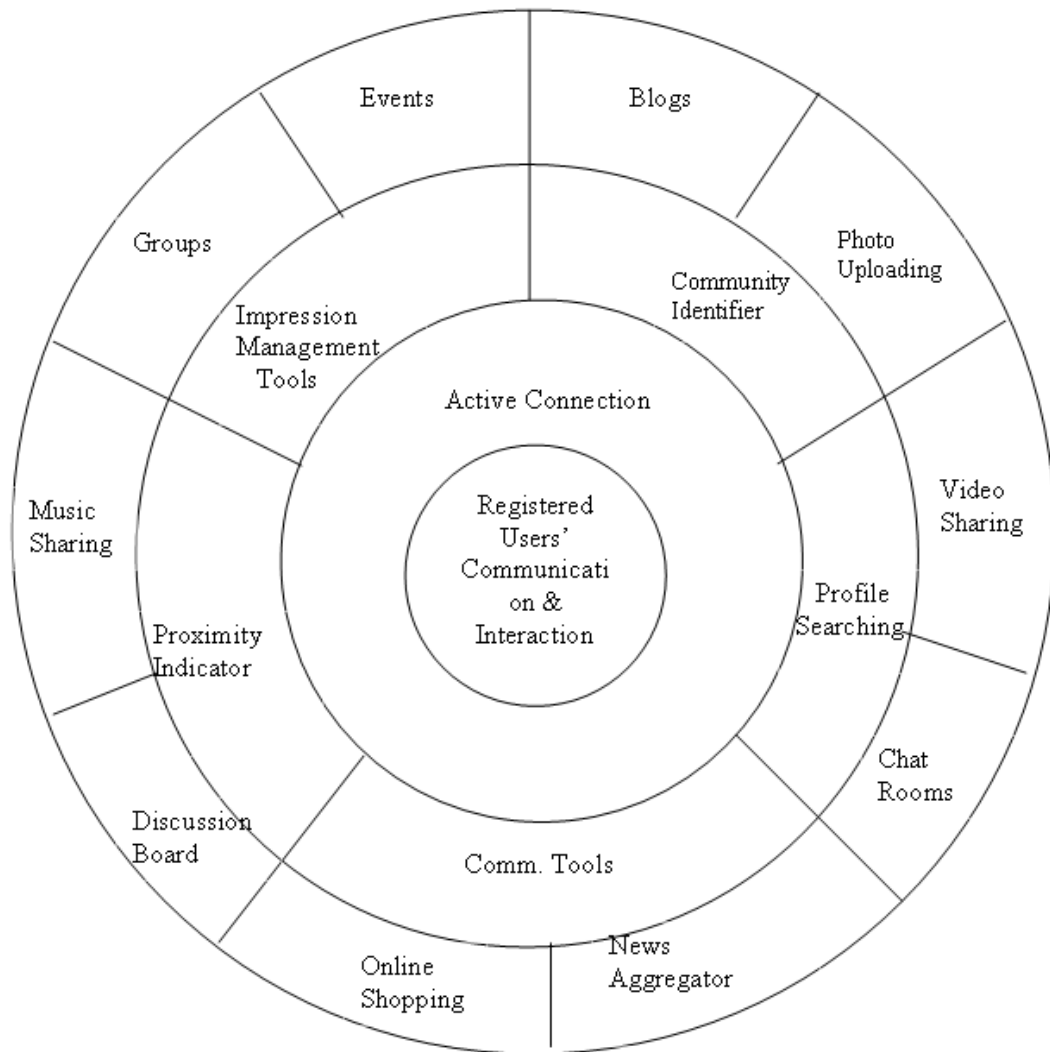
FIGURE 7.1: RealSpace Architecture

identifier to select different profiles suitable for different visitors. Because the visitors are again identified by an 8-digit identifier, the impression management tools can tell which profile is needed for which visitor.

Profile services should have their own characteristics on RealSpace system. The prototype aims to present different profiles of a user to his or her different friends, based on his or her initial setups. Users will have to configure their profiles on accessibility. For instance, a user can have multiple profiles to be viewed by relatives, colleagues, schoolmates, neighbours and other friends. They will have set up different content for different profiles. Each of this profile is associated with the internal identifiers of visitors. These identifiers are uniquely assigned to users when they register with the system. If a visitor who has been associated to a particular profile of another user and wants to access another profile of that user, he or she only can achieve this if that user has his or her privacy settings to be visible to this visitor. Different profiles can also be viewed

based on the proximity index. This means that the selection of profiles can be based on how close a visitor to the profile owner based on the social network of *active links*. Anonymous users, for example, may not able to view any of these profiles because they are considered as not in the social network.

7.2.2 Separation of Storage and Exchange Model

An important principle of the system architecture is the separation of storage schema and exchange schema. RealSpace stores the data in the relational database, which means we will model the data in the database in relational schema. These data will be used locally so that it will not compromise the efficiency of the system in the context of the Semantic Web. This principle is based on the observation that XML/RDF database is less efficient than traditional database and many systems are reluctant to abandon the existing workable data model and take the risk to move to a new unproven schema. On the other hand, the framework places paramount importance on the standardisation of the exchange schema, which is characterised by XML-based RDF format and widely used controlled vocabulary. The standardised exchanging format is particularly important in tackling the issue of walled garden social network sites. It is true that many social network sites such as Facebook and MySpace have developed their own set of file formats for data import and export. However, the vocabulary and file syntax they use share little in common and therefore they are not interoperable with each other. Google, on the other hand, has developed a social network framework, called OpenSocial, hopes that all the social network sites can adopt the standard. Thus, it is important to separate the storage and exchange model. But we are cautious to use heavy and rigorous ontologies, believing that the usage of these ontologies in the preliminary phase of the SW would only complicate the already limited exchanges between heterogeneous systems, making the Semantic Web vision a Semantic Utopia. The separation of storage and exchange schema is the fundamental principle that distinguish the RealSpace system from other otherwise similar Semantic Web applications.

7.2.3 Permission Granting

Consider the scenario that a user wants to share his or her friends with another user or a social network system wants to share its data from another social network system with a third party system. If all the individuals, agents and computer systems are happy to share (or keep private) all the available data to all parties that have established relationships with them, then we do not need any permission granting mechanism. Unfortunately this is not the case in the real world. There is a need to develop different authorisation levels in order to grant different users of different rights in using different data. Such permission granting mechanism serves a balance' role in the Semantic Web where data can freely flux in and flow out from various channels.

Permission granting is important in controlling the privacy settings. Users will usually have to decide which visitors can access which part of his or her profiles and personal information. Visitors need users' permission to view the information of the users. Users should hold the power to grant permissions to visitors. In fact, in many social network sites such as MySpace and Facebook, users will have a certain degree of permission control. But these permission granting mechanism is still in its infancy and leaves a large space for further development. In RealSpace system, as shown below, we will make the permission granting a part of the algorithm that attempts to identify the local social network.

7.2.4   Utility Programs

Some utility programs are designed to provide essential functionalities to RealSpace. These include two parts: complex network calculator and a cache module. The network calculator is used to estimate the parameters for complex network model of social network. The cache module is used to improve the browsing speed of the website.

The social network based on *active links* as identified by the ActiveLink algorithm should conform to the real-world social network. This means it should have a decent level of approximation to the topological features of real network such as power-law degree distribution, clustering coefficients, and average short path. The RealSpace core algorithm is based on a trial-and-error method and therefore the resulting social network will be compared with the real-world social network each time to verify its effect. This is likely to require significant computing resources. Therefore, we design a complex network checker to cope with the computing issues. The module specialises on calculations of specific parameters such as power-law, clustering coefficients, assortativity, average short path and so on. It reuses the previous results whenever it can to save the time and space.

RealSpace will use different registration. Any user with an email address can have full registration with the system. Even users who do not have email addresses can also register fully with the system. In fact, we have intended to work on an email system within the RealSpace system such that it can provide full service of a typical email system. Validation of new registration should be simplified. Catcha is a necessary evil for preventing spammers and phishers from attacking the system by running a computer script. But besides these, other options should be simplified as much as possible. More details, if the user wish to fill in, can be filled in later after they have signed up with the system and start to use it. We believe the issues such as loose acquaintances, fakesters and trust can be better addressed by using dynamic network algorithm such as the *ActiveLink* algorithm. Setting up the policy for validating registration is not a long-term solution. These policies are simply a weakness of the algorithms that are not able to reflect the real-world social network.

Another important improvement of accessing speed is the introduction of a cache module to the system. RealSpace search engine, which is programmed in PHP, is a web-based system. Depending on the bandwith, the client-server model will incur a level of latency so it is better to cache the results when browsing. This is particularly significant when the RealSpace system has already dedicated a huge amount of computing resources to deal with the networking algorithms. RealSpace employs MemcacheD as its caching system. It is a distributed memory object caching system that is also used by Facebook. The module is estimated to boost the software's performance by over 20% and increase its memory efficiency over 30% if new functionality is added[1].

7.3 System Structure

Figure 7.2 illustrates major component modules of the system. Solid lines indicate the modules that have been built, while dashed lines indicate the modules that have yet to be materialised. A rectangular box shows the module will read data from the database. An oval box shows the module contains interactions between users so that read/write operations are both required to be done on the database. This is the implementation of the idea that dynamic activities between the users should be registered to the activity checker for updating the record. We propose that these activities include but are not restricted to private messaging, instant messaging, public wall postings, blog comments, photo and video comments and exchanging gifts. Diamond boxes are auxiliary units that aim to improve the performance of the major components. For example, to accelerate the speed of querying the database, we will have to index the entries of the table. This is done by the auxiliary unit. For many Web applications, cache is essential for improving the accessing speed. Therefore, auxiliary units will also provide a PHP cache, presumably MemPHP. The activity checker, in particular, is responsible for refreshing the real connections between the people. This part will contain an algorithm to calculate the topological factors of the online social network. The coloured modules are parts of the architecture, most of which have been discussed in the previous chapter. The grey parts are routine components that are either necessary to the system or provide extended applications. The system is designed such that individual component is to be loosely coupled with each other.

7.3.1 Database Schema

The database is the soul and heart of the system. We use MySQL for out database. Unlike data repository of Web search engine, which generates the data by crawling the Web, a database of social network site captures the data input by the users. An HTML document may just include creation time, headline, metadata and full text while the record

---

[1]http://developers.facebook.com/opensource.php
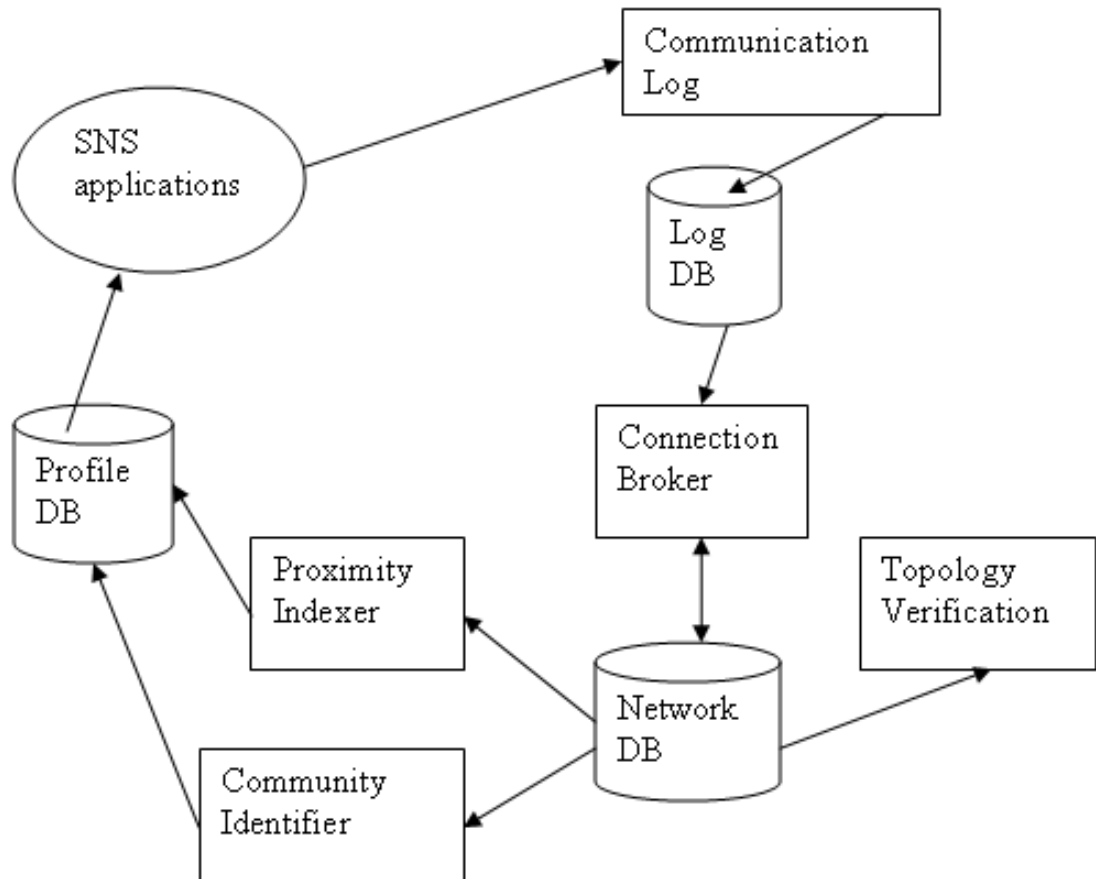
FIGURE 7.2: Major Component Modules

of a person will have many more dimensions of information, which can be very flexible. Thus, a detailed schema is necessary to provide rich descriptions of the data. In Figure 7.3, for example, the table *Users* includes $uid, firstname, surname, email, occupation$, etc. These information are either input by new members when they register or supplemented by existing members if they need to. There are two different types of data. The first type has enumerative values. These are gender and political views. Users can only select the value from a list of candidate options. The second type is number where the data is essentially numbers. This is convenient for doing mathematical calculation. The third type is generally text. Users can enter any information of their choice. The data will be used by almost all other modules and therefore we write a query interface that specifically inserts and retrieves the data from the table. The format of email will be checked before entering the table. Initially, this includes both the well form of an email (X@Y) and the valid form of the email which is eligible for registering. We drop the second criteria later. The design emphasises the *occupation* and *location*, as both social dimensions provide important cues for decentralised search. We restrict the universities and companies to a list which is maintained and constantly updated as the network grows.

The table of *buddyfriends* is to record the connections between people. These connections should not be confused with *active links* as we discussed in Chapter 5. The connections in the table of *buddyfriends* are purely raw connections as input by the users. They represent users' original activities and behaviours in the social network system. RealSpace system records these information in order to process them by using the ActiveLink algorithm. The field of *fuid* represents the user who initiates the friend request. Because this is a prototype, we do not intend to accommodate more than one billion users. Therefore, this field is an 8-digit positive number. The field *type* describes the category of friendship. These are a range of friends based on the nature of connections. For example, 0 may represent intimate friends, 1 for close friends, 2 for relatives, 3 for generally friends, 4 for colleagues, 5 for neighbours, etc. The field *reject* indicates how many times user *tuid* has rejected the invitation. The introduction of this parameter is particularly important in preventing spamming and phishing. User *fuid* is banned if his request has been rejected by the same user for more than three times. Once a user is banned, he or she may not be re-activated unless the event is reported to the website administrator. The table is critical in constructing the relationships in the network. The format appears to be a directional edge in our table and there will always be an even number of entries as the relationships in social network are undirectional. The directional tables is important for a large-scale social network site. When the network grows bigger, the database will grow correspondingly. If the connections are entwined with each other in the table, then it is very difficult to separate them and store them in different database in different server.

Users								
uid*	fname	sname	email	location	occupation	...		

buddyfriends								
ref*	fuid	tuid	type	reject				

priv_msgs								
msg_id*	subject	fuid	tuid	msg_time	msg_text	read	f_delete	t_delete

banned_users								
email*	timestamp							

FIGURE 7.3: Database Schema

7.3.2 Exhaustive Searcher

RealSpace's exhaustive searcher assists users to find people in the database. Members' information, such as their profiles, blogs, uploaded photos, videos and comments, have been transformed and saved in the RealSpace database. The raw data will be sorted and

indexed before they can be used. Among these applications is exhaustive searcher module. The module is essentially a search engine with sophisticated information services. There are two interfaces with the searcher: a general search interface and an elaborate search interface. The general search engine will be quick and easy to use. It is mainly for queries where the users know the names they are going to search. The elaborate search interface, as its name suggested, is much more sophisticated than the general one. It intends for queries on the database based on all the criteria available, subject to other users' privacy setting. The results of both search interfaces will be ranked based on the score combining the *proximity index* and *social connectivity*.

For more details, on the general search interface, a user can issue a query by the name of either the person or the organisation he or she belongs to. The searcher will then look up the table in the database for possible match. If the searcher finds an entry fully match the query term, then it will return it to the user. If the searcher find more than one entry that matches the criteria, it will use the ranking algorithms to rank the results. The ranking can be changed to be based on other criteria such as alphabetical order of surname, location and birthday, connections between the user and the people he or she wants to search.

On the elaborate search interface, as shown in Figure 7.4, one can make a query by specifying the detailed profile of the person. The criteria include but are not restricted to name, gender, birthday, political views, occupation, hobbies, working and education history and residence. The searcher will make an intersection operation on the queries and return the results. We do not use union operation on these criteria. The advantage of using intersection operation is that the returned list of people will only conform to all the conditions as described. At present, the result will be ranked alphabetically for our convenience. To accelerate the speed of searching, the elaborate search engine will introduce a cache service which stores the results that are searched very often. These are usually highly connected people and opinion leaders who occupy a central position in the social network. The application will also consider the alphabets of the names that is not one of the 26 English letters.

7.3.3   Validating Registered Users

The majority of social network sites have no restrictions as to who can join or when. The benefit of open registration is that users can have better chance to extend their networks. Such network will benefit from weak tie relationships greatly. MySpace is one of these examples. It opens to public registration in the first day it was launched to compete with other sites like Friendster. In contrast, Facebook did not open to public registration in the beginning when it was launched. Instead, it can only be registered with a designated university email account. As a result, Facebook grew slower than MySpace at the beginning of its development. The disadvantage of open registration,
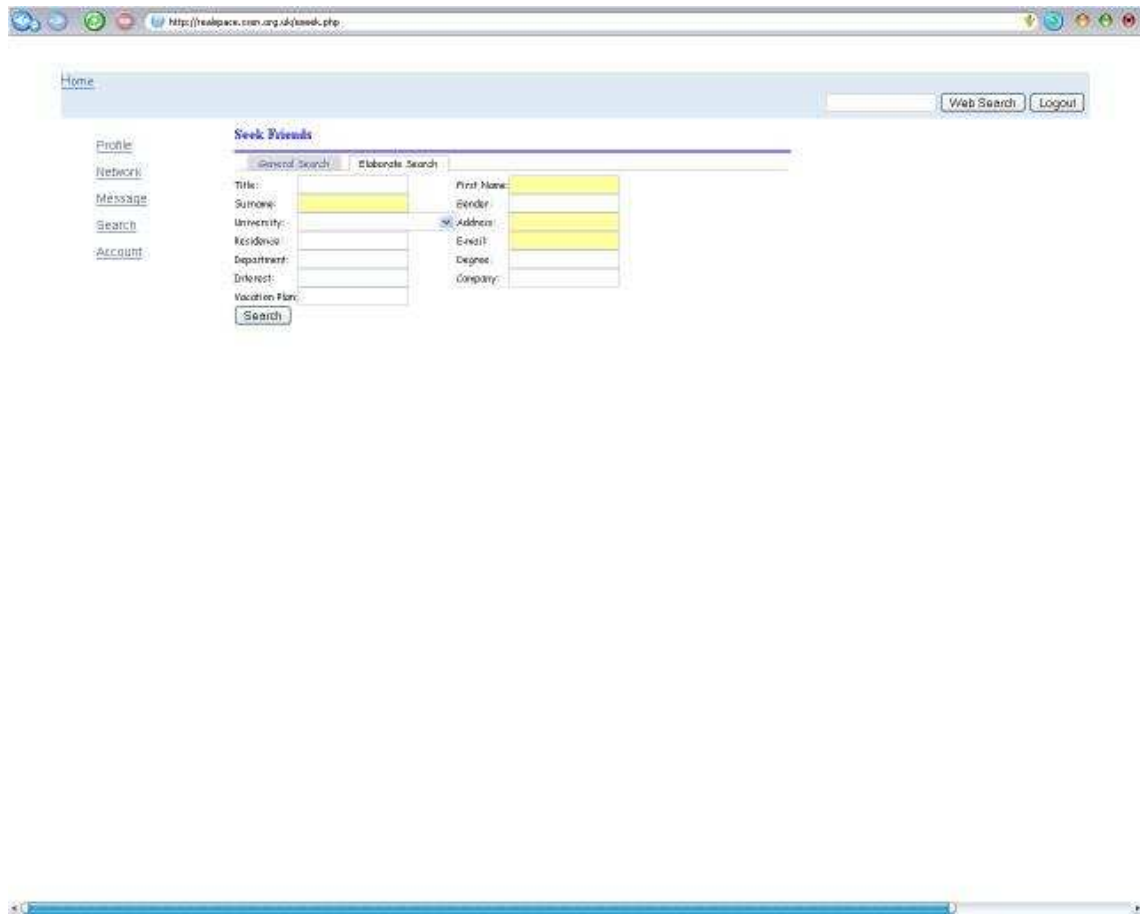
FIGURE 7.4: The Interface of Elaborate Search

however, is that there are less coherence and integration in the network. Users may feel less committed to the connections which are acquired through the websites. Because any one with a public email address can register with the site, spammers and phishers can gain full access to the website without much effort. Therefore, the social network feels not as safe as the one with designated email registration. Some network sites require a certain form of identifier or friends' invitation. Orkut and Facebook were examples of these kinds, though that requirement is now abolished due to the commercial interests. On these sites, fewer members would register in the beginning and the number of users may grow much slower than that of the open sites. However, there are more trust in the network as they mirror the real connections of the registered users. They might also reduce significant amount of loose acquaintances and fakesters. Due to the benefits of "open culture" in social networks, both Orkut and Facebook open their registration to general public. The change of the policy boosts user base but incurs problems that damage the reputation of the sites. It is unlikely for these sites to overcome the issues effectively as they use static model of social network.

RealSpace will use open registration. Any user with an email address can have full registration with the system. Even users who do not have email address can also register

fully with the system. In fact, we have intended to work on an email system within the RealSpace system such that it can provide full service of a typical email system. Validation of new registration should be simplified. Catcha is a necessary evil for preventing spammers and phishers from attacking the system by running a computer script. But besides these, other options should be simplified as much as possible. More details, if the user wish to fill in, can be filled in later after they have signed up with the system and start to use it. We believe the issues such as loose acquaintances, fakesters and trust can be better addressed by using dynamic network algorithm. Setting up the policy for validating registration is not a long-term solution. These policies are simply weakness of the algorithms that are not able to reflect the real-world social network.

7.3.4 Flexibility of Information Control

Users have all the rights to control their personal information. They should be able to decide what information to be revealed to whom. The information include subjective data such as the profiles users fill in by themselves, together with the objective data such as the number of active contacts and social connectivity which are calculated by the system. For audience, it could be different individuals or different groups of individuals. Flexibility should also be extended to outside the network if users would like to share their information with unregistered users. Many network sites provide privacy settings for users to control the information flow. However, commercial network sites have tendency to maximise the number of registration by displaying as more information about the existing members as possible. Therefore, the default setting has been usually revealed a significant amount of personal information about the users. These might benefit the users when the network is small and relationships are genuine. The revelation of information will come against the users when more users join the network and are able to access the information which otherwise is not intended to share with strangers.

In RealSpace, basic information such as nickname and location will display by default. All other information is not disclosed unless it is told so by the users.

7.3.5 Reputation and Trust

As mentioned in the earlier chapter, social network sites provide a trust layer on top of the Web platform. This trust layer is the social network of people's real-world relationships. However, while it appears intuitive to users, it lacks sophisticated analysis of the social network, particularly when it grows bigger in size. On SNSs like Facebook and MySpace, the number of registered users can well exceed one million. How can we identify the reputation of other users in a large-scale social network?

We argued in chapter 5 that proximity is an important factor in determining whether we trust one another. The assumption of this mechanism is that people trust members who are their direct contacts. The trust will propagate through the chain of social network. Therefore, the closer we are to a person, the more we trust him/her. The proximity index, which measures the closeness between any designated user and any other users in the social graph, can reflect the psychological distance in the landscape of trust.

Another element that contributes to trust is reputation. The higher reputation a user has, the more we trust him/her. On auction site such as eBay, user's reputation is usually indicated by a rating that is based on transaction. In social network, user's reputation can be measured by various prestige and analysis based on the structure of the network. The analysis technique of betweenness, for example, is a centrality measure that calculates the extent to which a node is directly connected only to those other nodes that are not directly connected to each other. Another technique, eigenvector centrality, measures the importance of a user by assigning a relative score to each user based on how many users he or she connects to and how influential they are. Our trust measure, based on proximity index, will take consideration on this centrality analysis. This parameter will have share the part of its weight of the final score of trust.

The large-scale social network differs from a simple network with less than thousands of users in that there can be multiple hubs and centres with similar degree of connections. These are important users who can affect other users in their circle of influence, but play a less important role in others' territories. On SNSs such as Facebook, for example, these territories are groups of different interests and purposes. These groups can also be formed based on geography. Each group may have its own active members that affect the group activities and organisation. But these opinion leaders may have less voice in other groups of which they are not a member. Thus, we will take into account the group structure of social network when integrating centrality analysis into reputation rating.

Finally, the groups created by users and joined by other users on a voluntary base, may or may not reflect the actual connections of members in the group. In a group such as student group from the University of Southampton with about one thousand members, because the users of the group normally study together in the same university and live in a relatively small city, they may well be connected to each other. In contrast, in a group such as London with up to one million users, there could be hundreds of fragmented subgroups with members from in- and outside London and from different walk of life. When we look at the scope of centrality analysis, we are not going to use these naturally formed groups, instead, we will utilise the community structure based on the connections between users. The communities are identified by the algorithm as mentioned on chapter 5. It will divide the social network into various communities based on how people are more closely connected to the part of users and less connected to anther part. The proximity index, combined with the centrality analysis based on community structure, will finally form the index of reputation and trust.

7.4 System Features

The *RealSpace* social network system introduces three features for social networking: *Social Connectivity*, *Proximity Index* and *Community Structure Detection*. All of the three algorithms are based on the active links. Social connectivity measures the importance of individual in the social network. Proximity index indicates the closeness between a user and every other users. Community structure detection can identify the groups of which members are more closely connected with each other than the rest of the network.

7.4.1 Social Connectivity

Most social network sites provide search function for members to find people. The search algorithm is typically an exhaustive search which retrieves all the items with criteria specified by the user. To do so, the algorithm needs to first index all the people in the database. When a query is issued, it looks up the table to locate the people related to the query. Since there are usually thousands of results returned, some kind of ranking mechanism is employed to sort the results. A primitive ranking algorithm is to rank by surname, as currently used in Facebook, but this algorithm is usually too naive to have any effect on the ranking. Another strategy that is currently used by some sites is rating-based ranking. However, the algorithm is easily subject to abuse by users as over-rating or under-rating. This is true particularly when the users see the benefits of doing so. Unfortunately people do benefit from such activities. The *preferential attachment*, as described in BA model, indicates that people tend to make friends proportional to the targeted individuals' degree. Thus, better connections will attract more friends. Therefore, we develop a ranking algorithm based on social connectivity to improve the search quality.

In a network $\mathcal{N}(\text{N,T})$, the social connectivity $C(u)$ for the person u is defined as follows:

$$C(u) = \sum_{i \in N_{\mathcal{N}}(v)} P(i)C(v_i) \tag{7.1}$$

where v_i is the ith active contact of u and $P(i)$ is the weight of the connection between u and v.

Social connectivity is essentially eigenvector centrality using active connection. In SNA, eigenvector centrality has long been used to signify the importance of a node in the network[120]. The observation that higher social connectivity will have higher degrees is also corresponding to situation of *preferential attachment*. As the value is based on active contact, socialising footprints that may contribute to the connectivity will be disregarded. Social connectivity can be therefore used as an indicator for search ranking.

Compared with ratings system where reputation is manually rated, our metric is more robust, objective and effective.

7.4.2   Proximity Index

Many social networks provide an indicator called *mutual friends*. One can estimate how close he or she is to a stranger by looking at how many mutual friends they share with. However, the indicator does not go beyond two degrees. As the network grows, there may be more and more people sharing the same number of mutual friends. In addition, it would be useful to tell which one is closer when both share the same number of mutual friends. The index is particularly important when one has to rank the users or compare different users. Let S be the set of all members of an SNS, we introduce our algorithm as follows:

Algorithm 5 Calculate Proximity Index

1: **for** Each node v in S **do**
2:      Apply Breadth-first Search to $S \Rightarrow S_1, S_2, S_3, ...$
3:      Let $S_0 = \{v\}$
4:      Apply Eigenvector Centrality to $S_n \cup S_{n+1}(n \geq 0) \Rightarrow E_1, E_2, E_3, ...$
5:      $V_n \leftarrow n.E_n(n > 0)$
6: **end for**

Starting at v, it first applies BFS to the whole set and assigns the people with the same degree into the same set. For example, S_1 contains the members who directly connect to v while S_2 represents the set of friends of friends. Arranged in increasing order, the sets are evaluated by using eigenvector centrality. It first calculates the eigenvalues of the S_1, then S_2 and so forth. The eigenvalue is an integer. Finally, we generate the proximity index by combining the degree of separation and eigenvalue. The resulting indices would be formatted in the form of 1.xx, 2.xx and so forth.

Proximity index is arguably more important to individual users than centrality analysis. Centrality analysis is a method to identify influential and prestige people in the social network, yet these people may not be quite relevant to you. Proximity index can identify the users who matter to you by calculating your network distance. While everyone sees the same centrality analysis, different people will see different proximity index. Furthermore, this proximity index may be applied to measure the trust of information published by other users and trace their credibility.

7.4.3   Community Structure Detection

It is common to find group applications on most social networks. Users who share similar interests, values and ideas may join the same group. It is not unusual to find groups with hundreds of thousands of members. The problem is, however, these members may not

have personal connections with each other. These people may stay in the same group simply under the name of some kind of interest. Users have to spend a fair amount of time to navigate through others' social network to identify what communities they belong to. Therefore, we provide an algorithm called community structure detection in our social network system. The module implements Newman's community algorithm[98]. The algorithm offers a relatively effective approach for finding and evaluating the community structure in the networks.

7.5 Applications and the Social Network

On the third and outmost layer of the system lies in the applications that are useful for the social network sites. These include essential utilities, impression management tools, network viewer and communication tools. These applications and programs enable rich activities and behaviours on the social network. The will add values to the RealSpace social network system. Lessons should be learned from the earlier development of SixDegrees which managed to attract users to register with the site but provided few services and programs. Users are quickly fed up with the site and left. Another important functions for these applications and services are communication channels. The key algorithm of RealSpace social network system, ActiveLink, is based on the observed information that are exchanged between the users involved. Therefore, it is important to create channels and activities that facilitate various types of communication.

7.5.1 Impression Management Tools

We have pointed out the problem of universal profile and generic persona earlier in Chapter 3. Most social network sites will provide users profile services for them to present themselves. The profile services can be regarded as a type of impression management tools. After registration, users will normally be asked to fill in their personal profiles. The information required for the profiles include but are not restricted to name, location, hometown, work and education history. The profile can then be viewed by the users, subject to the profile owner's privacy setting. The problem of universal profile is that user can only maintain one profile. Therefore, the content as seen by both their employers and parents will be the same. This will usually cause some social embarrassment and social dramas. To solve the problem, we propose a RealSpace impression management program. The key data structure of the RealSpace social network system is its social network based on *active links*. Every user is treated as a node in the network and is identified by an 8-digit identifier. By using minimum representation of social network users we separate the underlying data structure from the applications. Therefore, we can provide multiple profiles to a user. This is achieved by allowing users with 8-digit identifier to automatically present different profiles according to the profiles of different

visitors. Because the visitors are again identified by an 8-digit identifier, the impression management tools can tell which profile is needed for which visitor.

RealSpace social network system aims to present different profiles of a user to his or her different friends, based on his or her initial configuration. For instance, a user can have multiple profiles for relatives, colleagues, schoolmates, neighbours and other friends. They will have to set up different content for different profiles. Each of these profiles is associated with the internal identifiers of visitors. These identifiers are uniquely assigned to users when they register with the system. If a visitor who has been associated to a particular profile of another user and wants to access another profile of that user, he or she only can achieve this if that user has his or her privacy settings to be visible to this visitor. Different profiles can also be viewed based on the proximity index. This means, the selection of profiles can be based on how close a visitor to profile owner based on the social network of *active links*. Anonymous users, for example, may not able able to view any of these profiles because they are considered as not in the social network.

7.5.2   Network Viewer

Network viewer is an application to visualise the social graph as identified by the RealSpace algorithm. The visualisation of the social network makes it easier to navigate and explore. Network viewer would ideally require interaction between the social map and users. Thus, the technique of AJAX or programming language like Java will be most appropriate for this purpose. The viewer is based on the Java. The advantage is that it can have maximum interactivity and flexibility in designing. However, the drawback to Java-based network viewer for a Web-based social network system is that it consumes more computing resources and Internet bandwidth such that it can cause latency and sometimes severe delay. We only consider the RealSpace prototype as a software for proving and demonstrating the principles of the evolving social network model. Therefore the problem of software response and users' experience are not in our priority.

Social network viewer is capable of displaying two types of social network: the one based on *active link*, as we previously discussed; and the one based on the user's own ego-centric network. In many cases, user's ego-centric network is not exactly the same as the network calculated by *active link* algorithm. This is because users add and remove friends from private friend list, which may be viewed fully or partially by other users in the social network. When they maintain a list of large number of friends, they may not be able to exchange information with them quite often. Particularly if the number of contacts exceeds the Dunbar's number. Thus, these connections, which may be regarded by the users as genuine connections, are not exactly the *active links*, which is based on the continuous exchange of messages, gifts and comments on profiles, photos and videos. Still, these connections, while some of them may be weak ties and some of them may be

strong ties, can be important to the users. Therefore, RealSpace will include the design of a network viewer that aims to support both social graphs.

Social network based on *active links* is displayed in blue as the backbones of the social graph. Connections that are not found in this social network but are regarded as mutual connections by users will be displayed in grey and drawn in smaller line. By clicking the lines of connections, users can see the nature of connections if they have specified, for example, relatives, colleagues, schoolmates and other types of friendship. By clicking the nodes of the graph, users can view the brief profile of that particular user. The viewer will also support the community structure detection. The algorithm of *community structure detection* is discussed in the previous chapter. It aims to identify the groups of users who share similar interests and hobbies with similar social and cultural backgrounds. It recognises the groups of users based on the social network of *active links*. With this function, users can easily identify their ego-centric network and the network nearby. They will also be able to see how their friends belong to other groups and how they are connected with each other. The application can change the background of displaying colour.

7.5.3 Essential Utilities

We have some essential utility programs that provide behind-the-scene functionalities to the RealSpace social network system. These include two parts: complex network checker and PHP cache. The network checker is used to estimate the parameters for complex network model of social network. The cache module is used to improve the browsing speed of the website.

The core algorithm of RealSpace is to identify the *active links* and form the social network based on it. This resulting social graph should be conform to the real-world social network. This means it should have a decent level of approximation to the topological features of real network such as power-law degree distribution, clustering coefficients, and average short path. The RealSpace core algorithm is based on a trial-and-error method and therefore the resulting social network will be compared against the real-world social network each time to examine its effect. This is likely to require significant computing resources. Therefore, we design a complex network checker to cope with the computing issues. The module specialises on calculations of specific parameters such as power-law degree distribution, clustering coefficients, assortativity, average short path and so on. It reuses the previous results whenever it can to save time and space.

Another important improvement of accessing speed is the introduction of PHP cache to the system. RealSpace search engine, which is programmed in PHP, is a web-based system. Depending on the bandwith, the client-server model will incur a level of latency so it is better to cache the results when browsing. This is particularly significant when

the RealSpace system has already dedicated a huge amount of computing resources to deal with the networking algorithms. RealSpace employs MemcacheD as its caching system. It is a distributed memory object caching system that is also used by Facebook. The module is estimated to boost the software's performance by over 20% and increases its memory efficiency over 30% if new functionality is added[2].

7.5.4 Communication Tools

RealSpace is in essence a social networking tool. The availability and quality of communication tools are important for social network sites. As of this writing, many social network sites have already provided both asynchronous communication tools such as private messaging and synchronous communication such as instant messaging. Facebook, in particularly, offers a web-based instant chatting program. While these tools enrich users' communication experience, they leverage the power of social network, which can provide a layer of trust and security. We showcase an anti-spam email system in RealSpace. It will demonstrate the principle of how to use the social network as a layer of trust and security.

Email is usually regarded as the killer application of the Internet. It had been used in some form even before the development of the Internet. The format of email includes header and body. There are several fields in the header: From, To, Subject and Date. Other common header files include Cc, Bcc, Received and Reply-To. Many social network sites provide private messaging which is essentially a simplified version of email. They can usually be only sent to other users of the same site. There are also header and body and sometimes attachment. A key difference between email and SNS messaging is the use of protocol. Email employs several Internet protocols such as POP3, SMTP and IMAP. Web-based email system also utilises HTTP. In contrast, SNS messaging only utilises HTTP and the flow of messages is only achieved in the same system. This suggests that while email can reach to broader audience in different network, SNS messaging can only be used in the same system.

Another important difference between email and SNS messaging is that while the first communication tool is that to send a message on social network site, one normally have already maintained a friend link to the users they want to communicate with. The advantage and disadvantages of these different configurations are obvious. With email one can communicate with virtually all the users on the Internet, but he or she may also receive a huge amount of spams. With social network messaging, one usually communicates with the people they already know and is restricted to ego-centric network, they are not able to reach broader audience. RealSpace aims to take advantage of both communication methods while overcoming their respective disadvantages. This requires a combination of email applications and social network infrastructure.

---

[2]http://developers.facebook.com/opensource.php

RealSpace will first introduce email-specific protocols to social network private messaging. These protocols, such as those have been identified previously, will transform social network messaging service to a functioning web-based email system. This makes the RealSpace messaging service reach outside the system to many different systems which also support email protocols. Th email will be associated with a unique identifier from the RealSpace system unless they have been associated with another identifier. Therefore, users who do not register with RealSpace but communicate with users of the system will be regarded as peripheral users and will equally be assigned an identifier. If there exists information exchanging between these users and the registered users, then they will form *active links* with the registered users and become part of the RealSpace social network. Noted that these people will not appear on the social network site. For example, they will not appear on users' profile. But they are treated internally as a part of the social graph. Thus, we have a social graph of both registered members and unregistered users who have connections with the registered users in the RealSpace system. The RealSpace network is the core of the anti-spam email system.

Each user will have his or her own ego-centric network, as well as a part of the RealSpace global network that he or she is able to navigate. We may call this the user's local social network. As more users are aware of their privacy settings and take action to change their privacy preferences, different users belong to different parts of the RealSpace global network. In fact, every user may end up with his or her own local social network which is quite different from one another's. The local social network is the combination of a user's ego-centric social network and a part of the RealSpace global social network. On many social network sites such as MySpace and Facebook, the local social network could be hugely inflated because of friendship inflation. Users may only recognise their ego-centric networks. RealSpace local social network provides the key information for identifying the spams.

A typical email system may have several folders: inbox, sent mail, drafts and spams. The anti-spam email system based on social network system will add several extra folders: first degree mail, second degree mail, third degree mail, fourth degree mail, fifth degree mail, sixth degree mail and outside six degree. As their names suggest, an email from a friend who is directly connected to the user will go to his or her first degree mail folder. If the email sender is not presented in the user's local social network, then it will go to outside six degree mail folder. The value of the degree is based on the shortest path between the user and mail sender in the user's local social network. In fact, the initial categorisation of emails is based on the *proximity index*, an algorithm previously discussed in Chapter 5. The arrangement of emails based on this particularly categorisation aims to improve the efficiency of viewing emails. This is justified by the intuition that emails from our friends usually are more important to us.

For those mail senders who are outside the six degrees, if they send out emails that contain unsolicited contents, then they are likely to be treated as spammers. Similar

rules will be applied to detect spammers who are far away from the receivers in their local social network. Users may re-organise the initial categorisation by moving the mail senders from one folder to another. For example, if he considers the mail sender in the fourth degree folder is important to him, then he may move the sender to the first degree folder. Once the user make the move, the mail sender that has been moved will stay in the designated folder until the user changes it again. The local social network will be changed based on the change of the first degree folder.

7.5.5 People Search

RealSpace's people search function help users to find other members in the database. Members' information, such as their profiles, blogs, uploaded photos, videos and comments, have been transformed and saved in the RealSpace database. The raw data will be sorted and indexed before it can be used. Among these applications is people search module. The module is essentially a search engine with fine-grained searching criteria. There are two interfaces with the searcher: the first search interface and the second search interface. The first search engine will be quick and easy to use. It is mainly for queries where the users know the names they are going to search. The second search engine, as its name suggested, is much more sophisticated than the general one. It intends for a search on the database based on all the criteria available, subject to individual user's privacy setting. The results of both search engines will be ranked based on the score combining the *proximity index* and centrality analysis.

For more details, for the first search interface, a user can issue a query by name of either the person or the organisation he or she belongs to. The searcher will then look up the table in the database for possible match. If the searcher finds an entry fully match the query term, then it will return it to the user. If the searcher find more than one entry that match the criteria, it will use the ranking algorithms to rank the result. The ranking can be changed to be based on other criterion such as alphabetical order of surname, location, birthday, connections between the user and the people he or she wants to search.

For the second search interface, one can make a query by specifying the details of the person. The criterion include but are not restricted to name, gender, birthday, political view, occupation, hobbies, working and education history and residence. The searcher will make an intersection operation on the search and return the results. We do not use union operation on this criteria. The advantage of using intersection operation is that the returned list of people will only conform to all the conditions as specified. At present, the result will be ranked alphabetically for our convenience. To accelerate the speed of searching, the elaborate search engine will introduce a cache service which stores the results that are searched very often. These are usually highly connected people and

opinion leaders who occupy central positions in the social network. The application will also consider the alphabets of the names that is not in the 26 English letters.

7.6 Summary

In this chapter we discussed the detailed implementation of RealSpace social network system. RealSpace system has four layers of structure: the core layer is the communication and interaction between registered users that the system intends to capture; the second layer is the abstraction of the relationships using evolving social network model; the third layer is essential utilities of the system, such as impression management tools (profile editors), proximity indicator and community identifier; the outmost layer is the various applications such as blogging, video and music sharing. The system have several important applications: impression management tools, essential utilities and communication tools. We introduced the algorithms of social connectivity, proximity index and community structure detection. For communication tools, in particular, we discuss the anti-spam email system which employs the social network based on *active links*.

Chapter 8

Future Work and Conclusion

8.1 Conclusion

In this thesis, we have analysed the problems that challenge today's social network sites. These problems include friendship inflation, universal personas, privacy concerns, etc. In particular, friendship inflation, which is caused by the clash between network publicity and individual privacy, triggered by the technique of *static link*, has become one of the major issues in today's social network sites. Publicity and privacy are two fundamental forces that drive the development of online social network. Without publicity, users can not browse others' social network. Without privacy, users risk to expose themselves to strangers and spammers. A balance should be carefully negotiated between system designers and users. Unfortunately, the technologies currently employed by most SNSs such as *static link* and universal profile brutally damage the delicate balance. While the problem of universal profile may be remedied by providing multiple profiles and communication channels, there are no easy solutions to friendship inflation, as we show before. However, friendship inflation causes far more damages to the integrity and usefulness of the social network.

A hyperfriendship network model was proposed as a theoretical framework to describe the evolution of online social network. By preserving the rewiring edges the model shows how the online social network is developing. The topological differences include fast ageing and dissortative mixing. Then, we discuss the issues incurred from friendship inflation. The problems include unreliable connections, undiscernible hubs, lack of peer pressure, spamming and phishing, inaccuracy of network algorithms and information overload. We argue that friendship inflation will eventually lead to the decline of social network sites. To support the argument, we cite the case of the rise and fall of Friendster, MySpace and Facebook.

To tackle the problem, we proposed RealSpace, a social network system based on evolving social network model. The main objective of this work was to overcome the friendship

inflation problem by introducing active link, which is based on complex network theory. To achieve this objective, we first define the concept of *continuous reciprocity*, which can be seen on both direct and indirect communication. Based on Dunbar number theory, we impose a connection cap, which is the maximum number of connections the average users have. The connection decaying model, based on the forgetting curve of human beings' brain, was used to detect the obsolete connections.

The novelty of the work was the integration of topological features of complex network to the evolving social network. There are two main characteristics of people's social network: preferential attachment and assortativity. Both features guarantee the short paths between any two users in the network and therefore accelerate the dissemination of information and knowledge. They also make the network more robust and resilient. To achieve this, we allow users who have already maintained a higher-than-average number of connections to make more connections with less effort than average. The algorithm will also prevent these highly influential users from being abusing the power of connection by reserving the upper connections only taking place between these users, whose number is a small percentage of the whole population of the network. We gave a detailed description of what affects active connection and how it works. We explained the algorithm for evaluating active connection. Finally, we introduced the system features of social connectivity, proximity index and community structure detection.

8.2 Further Work

Besides this work on social network systems, a number of area of interest came to our attention which we were not able to further develop or study due to time constraints. In this section, we summarise the areas which we consider to be worthy of future research and outline a possible path for the future development of the software discussed in this thesis. These include complex network theory, managing the range of connection strength, improving the ranking of decentralised search algorithm, implementing the remaining modules that have not been realised in the prototype, and social network portability.

8.2.1 Complex Network Theory

Our goal is to support meaningful social networks. The idea is conceived according to our model which predicts the problem of growth constraint in many social network sites. Our model supplements the BA model with the key element of Kleinberg's model, that is, *long-range* shortcuts in power-law degree distribution. The model is used to explain the growth of social network sites qualitatively rather than quantitatively. A detailed computer simulation should be done to make the model more convincing. Furthermore,

a rigorous mathematical proof that *long-range* rewiring in BA model can exhibit the same topological features of complex network, such as small-world effect, large clustering coefficient and power law degree distribution, should be in the future research agenda. Current research has suggested that the clustering coefficient with BA model, though relatively large, is still not independent of the network size. We consider that since the BA model only takes into account the factors of *growth* and *preferential attachment*, the *long-range* rewiring in a power distribution fashion should provide some clues to overcome the weakness of the model.

When many users have more connections than they actually do, the topology of the network will increasingly diverge from real-world social network. We would like to develop a more sophisticated model to simulate the growth and evolution of the cumulative network.

The future model should be based on BA network as discussed in the second chapter. It has been observed that both conditions in the original model, *growth* and *preferential attachment*, apply to social network sites. In addition, there are three open questions to the model:

(a) In BA model, the exponent $\alpha=3$, but in real network, the number is between 2 and 3. We use 2.3, which is the measure for film actor collaboration network based on Internet Movie Database (IMDb). What will this parameter be if we combine both of the model?
(b) BA model does not specify the value of m, the average degree of the network. Dunbar's number suggests that people are capable to keep regular contact with about 150 friends. The number can be interpreted as the lower bound number of links one can have. Therefore the value of m, which is the number of friends that people claim to have, should be no less than Dunbar's number. For our convenience, m is set to be 150. How this will be changed if we combine both of the model?
(c) Individuals will make new acquaintances and forget old links after joining the network. This is called edge rewiring. BA model does not take into account the effect of internal edge rewiring. We assume in our model that every node will rewire his m edges to other nodes with probability p_r proportional to d^{-r}, where d is the social distance (described in chapter 2) between them and r is an adjustable constant. This condition will only be used qualitatively in our model. However, how will the edge rewiring change if we combine the models?

The remaining part of the population are able to maintain stable contact with more than 150 people. This is true regardless of the size of the network as it is scale-free. Notice here the notion of friends at least includes family members, neighbours whom you know and people whom you have worked or studied with for some time. So far, empirical data shows none of the social network sites gain the percentage of 44.78% or above, indicating that people have not yet fully moved their real-world relationships online. However, as the social network sites grow rapidly in the recent years, we would

expect the percentage will approach to that of the real-world network in a short period. Condition (c) suggests that people will "rewire" the friend links if they could not afford to keep regular contact with them, thus leaving a long trail of socialising footprints. In cumulative networks, the obsolete connections will not disappear automatically, which is contrast to the real social networks where old relationships will decay gradually if people do not keep regular contact with each other. Given the change of the model, how will this be affected?

8.2.2 Future System Development

Four pieces of work have been identified to complete the research. First, we need to merge the gap between Kleinberg's lattice model and the BA model. This will provide better theoretical framework for our system. Second, loose acquaintances can be distinguished from close friends in our system. But there are no effective management on acquaintances, who may make great contribution to the network due to weak tie effect. Thus, better categorisation of acquaintances should be developed to support the network. Third, the decentralised search algorithm simply utilises two or three social dimensions, in conjunction with closeness measure. Finally, we need to finish the remaining parts of the system according to our design.

8.2.2.1 Managing Connection Strength

So far, our system can only determine two types of relationships: acquaintances and close friends. The system ignores the loose acquaintances as *social footprints*. While the amount of close friends is small, the number of acquaintances is huge. Further, these acquaintances represent a whole range of social dimensions different from one's close network. One of the strength of social network sites is to retain history of all these connections, allowing users to accumulate and utilise the contact resources without memorising them. Thus, a useful social network site should not only identify the social footprints automatically but also take full advantage of them. Therefore, we would like to examine the range of connection strength. The focus is switched from nodes to ties. Inspired by the formula of learning curve, we are particularly interested in testing the hypothesis that the connection strength of social network displays a power law degree distribution. The hypothesis should be further scrutinised against data from social network sites and should be consistent with existing models.

We mainly focus on active connection, which refers to the connection between users who often exchange and share information. The methods for exchanging messages include both direct communications such as private messaging and instant messaging and indirect communications such as public wall posts, blog commenting, photo and video commenting and virtual gift exchanging. Instead of assuming a zero-cost establishment

of connection, it levies certain amount of communication effort to maintain the connection. The idea will be translated into the practice that the system will no longer employ the *static link* that takes a few clicks to befriend with one another, instead, it will look at how users interact with others whom they have added as friends and only the presence of continuous communication will signal connection. Many social network systems which recognise the weakness of the *static link* may devise a new algorithm for social network connection based on users' behaviours and activities. However, they rarely consider the role of social capital in determining the number of connections each user can acquire. Active connection is designed to be consistent with some topological features as found in the social network, such as *Preferential Attachment* and *Assortativity*. It will also take into account the factors of ageing and cognitive limit of human beings' brain. To illustrate the model, we compare the network of active connections with representative democracy model. Beyond the active connection, we would like to know more details about the acquaintance connections.

8.2.2.2 Decentralised Search

Our design of decentralised search algorithm brings transferrable social table and closeness to Watts' social distance model. We would expect the algorithm can yield better performance and is more reliable, yet a numerical simulation is still required to justify the prediction.

On the other hand, Watts' social dimension model claims that only two or three social dimensions are needed to achieve the construction of short paths and even lead to the best performance, in comparison with other choices. This may be true for the majority, but for the 20% of the population who have many more contacts than common people, the dimension number of two or three may be underestimated. We suggest a change on the model can be made on providing different choice based on the node degree (individuals' contact) to see if there is any improvement on the network navigation. None of the search algorithms we have reviewed so far consider the motivation issue when forwarding a message. The empirical observations suggests that friends who have closer relationships (strong ties) are more eager to help find the item of interests and pass the message to their friends more carefully.

Previously we develop a RealSpace exhaustive searcher, which can assist users to find other users in the social network. Members' information, such as their profiles, blogs, uploaded photos, videos and comments, have been transformed and saved in the RealSpace database. The raw data will be sorted and indexed before it can be used. Among these applications is exhaustive searcher module. The module is essentially a search engine with fine-grained searching criteria. There are two interfaces with the searcher: a general search interface and an elaborate search interface. The general search engine will be quick and easy to use. It is mainly for queries where the users know the names they

are going to search. The elaborate search engine, as its name suggested, is much more sophisticated than the general one. It intends for the search on the database based on all the criteria available, subject to individual user's privacy setting. The results of both search engines will be ranked based on the score combining the *proximity index* and centrality analysis.

We would like to improve the general search interface, where a user can issue a query by name of either the person or the organisation he or she belongs to. The searcher will then looks up the table in the database for possible match. If the searcher finds an entry fully match the query term, then it will return it to the user. If the searcher find more than one entry that match the criteria. It will use the ranking algorithms to rank the result. The ranking can be changed to be based on other criteria such as alphabetical order of surname, location, birthday, connections between the user and the people he or she want to search.

Another area we are currently looking at is the second search interface, as shown in Figure 7.4, one can make a query by specifying the detailed information of the person. The criteria include but are not restricted to name, gender, birthday, political view, occupation, hobbies, working and education history and residence. The searcher will make an intersection operation on the search and return the results. We do not use union operation on this criteria. The advantage of using intersection operation is that the returned list of people will only conform to all the conditions as described. At present, the result will be ranked alphabetically for our convenience. To accelerate the speed of searching, the elaborate search engine will introduce a cache service which stores the results that are searched very often. These are usually highly connected people and opinion leaders who occupy a central position in the social network. The application will also consider the alphabets of the names that is not in the 26 English letters.

8.2.2.3   Implementing the Remaining Components

The RealSpace prototype has laid out a foundation for the future development, yet many programming still needed to complete the important parts of the design. These include the *activity checker*, *social connectivity* and *closeness* calculator and more importantly, *decentralised searcher*. We would expect the modules of *activitychecker*, *socialconnectivity* and *closeness* should be written in C/C++ to improve the efficiency.

We have already developed some essential utility programs that provide behind-the-scene functionalities to the RealSpace system. These include two parts: complex network checker and PHP cache. The network checker is used to estimate the parameters for complex network model of social network. The cache module is used to improve the browsing speed of the website.

Here, the core algorithm of RealSpace is to identify the *active links* and form the social network based on it. This resulting social network should conform to the real-world social network. This means it should have a decent level of approximation to the topological features of real network such as power-law degree distribution, clustering coefficients, average short path. The RealSpace core algorithm is based on a trial-and-error method and therefore the resulting social network will be compared with the real-world social network each time to verify its effect. This is likely to require significant computing resources. Therefore, we design a complex network checker to cope with the computing issues. The module specialises on calculations of specific parameters such as power-law, clustering coefficients, assortativity, average short path and so on. It reuses the previous results whenever it can to save the time and space. We still need to add more functionalities to this module.

Another area that need to be improved is the accessing speed. RealSpace search engine, which is programmed in PHP, is a web-based system. Depends on the bandwith, the client-server model will incur a level of latency so it is better to cache the results when browsing. This is particularly significant when the RealSpace system has already dedicated a huge amount of computing resources to deal with the networking algorithms. RealSpace employs MemcacheD as its caching system. It is a distributed memory object caching system that is also used by Facebook. The module is estimated to boost the software's performance by over 20% and increase its memory efficiency over 30% if new functionality is added[1]. However, this is still not yet sufficient for our system.

For the prototype, the social network viewer is capable of displaying two types of social network: the one based on *active links*, as we previously discussed; and the one based on the user's own ego-centric network. In many cases, user's ego-centric network is not exactly the same as the network generated from the ActiveLink algorithm. This is because users add and remove friends from private friend list, which may be viewed fully or partially by other users in the social network. When they maintain a list of large number of friends, they may not be able to exchange information with them quite often. Particularly if the number of contacts exceeds the Dunbar's number. Thus, these connections, which may be regarded by the users as genuine connections, are not exactly the *active links*, which is based on the continuous exchange of messages, virtual gifts and comments on profiles, photos and videos. Still, these connections, while some of them may be weak ties and some of them may be strong ties, can be important to the users. Therefore, we will look at methods to distinguish these ties.

RealSpace social network system has already maintained different profiles of a user to his or her different friends, based on his or her initial setups. Users will have to configure their profiles on accessibility. For instance, a user can have multiple profiles for relatives, colleagues, schoolmates, neighbours and other friends. They will have to set up different content for different profiles. Each of these profiles is associated with the

---

[1] http://developers.facebook.com/opensource.php

internal identifiers of visitors. These identifiers are uniquely assigned to users when they register with the system. If a visitor who has been associated to a particular profile of another user and wants to access another profile of that user, he or she only can achieve this if that user has his or her privacy settings to be visible to this visitor. Different profiles can also be viewed based on the proximity index. This means, the selection of profiles can be based on how close a visitor to the profile owner based on the social network of *active links*. Anonymous users, for example, may not able able to view any of these profiles because they are considered as not in the social network. But how can we prevent the anonymous users to view the users' profiles by using different registered IDs? After all, many social network sites are open to public and registering an account is fairly simple.

8.2.3   Reputation and Trust

Reputation and trust are essential to the success of RealSpace. People are free to publish information and share opinions on the Internet, yet they can easily disguise themselves by remaining anonymous. As the Web keeps growing in a very fast pace, there are a huge amount of websites of different types that will produce tons of information and stories, some of which may well be rumours. It is difficult to tell the rumours from facts in different expertise domains which the readers are not familiar with, not to mention the slightly exaggerated stories and deliberately bias views that appears more subtle and undiscernible. The online anonymity makes it difficult to hold people responsible for their activities and behaviours. This will cause many ramifications and problems such as spamming, malware, online security, trust and privacy concerns. Therefore, there have been proposals to argue for the establishing of social web based on the existing Web and Internet infrastructure. The idea is to bring trust and security to Internet by leveraging the peer-to-peer pressure on individuals. The accountability can be achieved with SNSs as public pressure can be formed due to the difficulty to remain anonymous, as users generally publicise the connections to their real-world friends. Friendster, with its social network reach of four degrees when it was launched, is one of the first dating sites to take advantage of the publicity of profiles and contacts to provide trust and security. Our prototype still needs to improve the reputation and trust ratings.

8.2.4   Social Data Portability

With the hundreds of SNSs on the Web and many more in different languages from different countries are coming, there are increasing concerns about the interoperability between these walled garden SNSs. If a user who has registered with Facebook but wants to access the social network on MySpace, he or she must create an account on MySpace and fill in all the details again and add friends, which has been done previously

on Facebook. Most social network sites allow people to have their data exported via some applications. But this is usually only restricted to profiles. The profiles were very simple in the first generation of social network sites, they become increasingly rich in description, thanks to the advancement of Web technologies and standards. A profile usually includes but is not restricted to name, birthday, current location, hometown, interests, education and work history. Some may also display information about their social networks, relationship status, contact methods, etc. Some of these options are enumerative, such as gender and political views. This means users can only select one from the pre-defined list. Many options can be filled in with limited number of text, some of which can even utilise the features of HTML and Javascript language. It is also commonplace for users to upload photos and videos to their profiles. Profiles usually employ WYSIWYG (What You See Is What You Get) method so that the editing effect can be seen immediately. Users can change the permission of the profiles that can be accessed by other members. However, if any part of the profiles can be accessed, they are viewed by all visitors to be the same content. It can not detect the visitors based on the nature of the connections such as parents and employers. Thus, the profile is fixed and universal on the social network sites. Users who have already maintained several accounts on different sites also face the problem of synchronisation. They are at pain to constantly update the information on different account in order to keep them relevant. Hence, it is important for different social networks to talk to each other.

As of this writing, SNSs do provide some solutions to this problem. Among the first move is MySpace's Data Availability program, which provides MySpace social network system to third party websites which want to integrate the social networking features. This follows by Google Friend Connect, Facebook Connect, etc. Each of these sites can offer their own social network systems to third party sites. While this extends the reach of social network, there are still no solutions for different social network systems to talk to each other. There have long been academic proposals to solve this problem. These include FOAF, Social Web's standards such as XDI, openID, Oauth, etc. The fundamental idea behind these methods is to decompose the social network system and grant the users more permissions to handle their own data. On the other hand, users will have more responsibility for their activities in the social network. However, given the commercial interest, security and privacy issues, it is yet to see whether SNSs will adopt these solutions and how far they can go.

Bibliography

[1] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *6th International Workshop on Privacy Enhancing Technologies*, volume 4258, pages 36–58, 2006.

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

[3] P. Adler and S. Kwon. Social capital: Prospects for a new concept. *Academy of Management Review*, 27:17–40, 2002.

[4] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. 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.

[5] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, (74):47–97, 2002.

[6] R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks. *Nature*, (406):378–382, 2000.

[7] A. Ankolekar, G. Szabo, Y. Luon, B. A. Huberman, D. Wilkinson, and F. Wu. Friendlee: A mobile application for your social life. In *Mobile HCI 09*, Bonn, Germany, 2009. ACM.

[8] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of 12 th International Conference on Knowledge Discovery in Data Mining*, pages 44–54, New York, NY, USA, 2006.

[9] A.-L. Barabasi. *Linked - How Everything is Connected to Everything Else and What it Means for Business, Science, and Everyday Life*. Plume Books, 2003.

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

[11] J. A. Bargh and K. Y. A. McKenna. The internet and social life. *Annual Review of Psychology*, 5:573–590, 2004.

[12] S. Barnes. A privacy paradox: Social networking in the united states. *First Monday*, 7:39–58, 1954.

[13] N. S. Baron. My best day: Presentation of self and social manipulation in facebook and instant messaging. In *Eighth International Conference, Association of Internet Researchers*, 2007.

[14] J. Baudrillard. *Simulations*. New York: Semiotex, 1983.

[15] P. Bourdieu. The forms of capital. *Handbook of theory and research for the sociology of education*, pages 241–258, 1986.

[16] D. Boyd. Friendster and publicly articulated social networks. In *Conference on Human Factors and Computing Systems*, Vienna, Austria, 2004.

[17] D. Boyd. Friends, friendsters, and myspace top 8: Writing community into being on social network sites. *First Monday*, 11(12), 2007.

[18] D. Boyd. None of this is real. *Structures of Participation in Digital Culture (ed. Joe Karaganis)*, pages 132–157, 2008.

[19] D. Boyd and N. B. Ellison. Social network sites: Definition, history, and scholarship. *Computer-Mediated Communication*, 1(13):11, 2007.

[20] D. Boyd and J. Heer. Profiles as conversation: Networked identity performance on friendster. *Proceedings of Thirty-Ninth Hawai'i International Conference on System Sciences*, 2006.

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

[22] J. Brown. Six degrees to nowhere. *Salon.com*, 1998.

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

[24] L. Charnigo and P. Barnett-Ellis. Checking out facebook.com: The impact of a digital trend on academic libraries. *Information Technology and Libraries*, 26(1): 23, 2007.

[25] W. Chung, R. Savell, J.-P. Schutt, and G. Cybenko. Identifying and tracking dynamic processes in social networks. In *Sensors, and Command, Control, Communications, and Intelligence (C3I)*, volume 6201 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, Jun 2006.

[26] T. Coenen, D. Kenis, and C. V. Damme. Knowledge sharing over social networking systems: Architecture, usage patterns and their application. volume 4277 of *On the Move Federated Workshop*, pages 189–198, 2006.

[27] J. S. Coleman. Social capital in the creation of human capital. *American Journal of Sociology*, 94:S95–S120, 1988.

[28] L. da F. Costa, N. Osvaldo Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, and L. E. C. da Rocha. Analyzing and modeling real-world phenomena with complex networks: A survey of applications, 2007.

[29] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances In Physics*, 56: 167, 2007.

[30] J. M. DiMicco and D. R. Millen. Identity management: multiple presentations of self in facebook. In *In GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, pages 383–386, 2007.

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

[32] P. Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20(1):80–82, 2005.

[33] J. Donath. Signals in social supernets. *JCMC*, 13(1), 2007.

[34] J. Donath and D. Boyd. Public displays of connection. *BT Technology Journal*, 22(4):71–82, 2004.

[35] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites. *Phys. Rev. E*, 62(2):1842–1845, Aug 2000.

[36] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51:1079, 2002.

[37] R. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–735, 1993.

[38] C. Dwyer. Digital relationships in the 'myspace' generation: Results from a qualitative study. In *Proceedings of the 40th Hawaii International Conference on System Sciences*, 2007.

[39] C. Dwyer, S. R. Hiltz, and K. Passerini. Digital relationships in the 'myspace' generation: Results from a qualitative study. In *Proceedings of AMCIS 2007*, 2007.

[40] H. Ebbinghaus. Memory: a contribution to experimental pychology. *Dover: New York*, 1885.

[41] N. Ellison, C. Lampe, and C. Steinfield. Spatially bounded online social networks and social capital: The role of facebook. *International Communications Association*, 2006.

[42] N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 2007.

[43] N. B. Ellison, C. Steinfield, and C. Lampe. Social network sites and society: Current trends and future possibilities. *Interactions Magazine*, 16(1), 2009.

[44] P. Erdos and A. Renyi. On random graphs. *Publicationes Mathematicae*, (6): 290–297, 1959.

[45] R. Ermecke, P. Mayrhofer, and S. Wagner. Agents of diffusion insights from a survey of facebook users. In *Proceedings of the Forty-second Hawaii International Conference on System Sciences (HICSS-2007)*, Los Alamitos, CA, 2009.

[46] D. Fono and K. Raynes-Goldie. Hyperfriendship and beyond: Friends and social norms on livejournal. in m. consalvo & c. haythornthwaite (eds.). *Internet Research Annual Volume 4: Selected Papers from the AOIR Conference (pp. 91-103)*, 2006.

[47] S. Fox. Trust and privacy online: why americans want to rewrite the rules. Technical report, The Pew Internet & American Life Project, Wahington DC, Jun 2000.

[48] H. Gardner. *Frames of mind: the theory of multiple intelligences.* New York: Basic Books, 1983.

[49] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*, 2002.

[50] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference.* Little Brown, 2000.

[51] E. Goffman. *The presentation of self in everyday life.* Garden City, NY: Doubleday and Co., 1959.

[52] Jennifer Golbeck and James Hendler. Inferring binary trust relationships in web-based social networks. *ACM Trans. Internet Technol.*, 6(4):497–529, 2006.

[53] S. A. Golder, D. Wilkinson, and B. A. Huberma. Rhythms of social interaction: Messaging within a massive online network. In *3rd International Conference on Communities and Technologies (CT2007)*, East Lansing, MI, 2007.

[54] V. Gomez, A. Kaltenbrunner, and V. Lopez. Statistical analysis of the social network and discussion threads in slashdot. In *WWW2008*, 2008.

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

[56] M. Granovetter. The strength of weak ties: Revisited. *Sociological Theory*, 1: 201–33, 1983.

[57] R. Gross and A. Acquisti. Information revelation and pivacy in online social networks. In *Workshop on Privacy in the Electronic Society*, Alexandria, VA, 2005.

[58] J. Guare. Six degrees of separation: A play. *Vintage Books, New York*, 1990.

[59] C. Haythornthwaite. The strength and the impact of new media. In *HICSS '01: Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 1*, page 1019, Washington, DC, USA, 2001. IEEE Computer Society.

[60] J. F. Helliwell and R. D. Putnam. The social context of well-being. *Philosophical Transactions of the Royal Society*, 359:1435–1446, 2004.

[61] James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. Web science: an interdisciplinary approach to understanding the web. *Commun. ACM*, 51(7):60–69, 2008. ISSN 0001-0782.

[62] S. C. Herring, J. C. Paolillo, I. Ramos Vielba, I. Kouper, E. Wright, S. Stoerger, L. A. Scheidt, and B. Clark. Language networks on livejournal. In *Proceedings of the Fortieth Hawai'i International Conference on System Sciences*, Los Alamitos, CA: IEEE Press, 2007.

[63] P. Holme, C. R. Edling, and F. Liljeros. Structure and time-evolution of an internet dating community. *Social Networks*, 26:155, 2004.

[64] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), 2009.

[65] N. Humphrey. *The Social Function of Intellect*. Cambridge University Press, 1976.

[66] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[67] T. Jagatic, N. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Communications of the ACM*, 5:94–100, 2007.

[68] E. M. Jin, M. Girvan, and M. E. J. Newman. The structure of growing social networks. Working Papers 01-06-032, Santa Fe Institute, Jun 2001.

[69] B. Johnson. Twenty-eight people ask hugh macleod to be their friend each day. what's so special about him? *The Guardian*, December 2007.

[70] D. Johnson, S. Crawford, and J. G. Palfrey. The accountable net: Peer production of internet governance. *Virginia Journal of Law and Technology*, 9(9), 2004.

[71] A. N. Joinson. Looking at, looking up or keeping up with people?: motives and use of facebook. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1027–1036, New York, NY, USA, 2008. ACM.

[72] K. Jordan, J. Hauser, and S. Foster. The augmented social network: Building identity and trust into the next-generation internet. *First Mondy*, 8(8), 2003.

[73] S. Jurvetson. What exactly is viral marketing? *Red Herring*, 78:110–112, 2000.

[74] T. P. Kiehne. Social networking systems: History, critique, and knowledge management potentials. *Student Paper*, 2004.

[75] P. D. Killworth and H. R. Bernard. The reverse small world experiment. *Social Networks*, (1):159–192, 1978.

[76] J. M. Kleinberg. The small-world phenomenon: an algorithmic perspective. in Proceedings of the 32nd Annual ACM Symposium on Theory of Computing, 2000.

[77] J. M. Kleinberg. Complex networks and decentralized search algorithms. Proceedings of the International Congress of Mathematicians (ICM), 2006.

[78] V. E. Krebs. Uncloaking terrorist networks. *First Monday*, 7(4), 2002.

[79] C. Lampe, N. Ellison, and C. Steinfield. A face(book) in the crowd: social searching vs. social browsing. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 167–170, New York, NY, USA, 2006.

[80] J. Laraqui. Activity based interfaces in online social networks. *Thesis (M. Eng.)– Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science*, 2007.

[81] S. H. Lee, P. J. Kim, and H. Jeong. Statistical properties of sampled networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 2006.

[82] A. Lenhart and M. Madden. Teens, privacy, & online social networks. *Pew Internet and American Life Project Report*, 2007.

[83] Y. I. Leon-Suematsu and K. Yuta. A framework for fast community extraction of large-scale networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1215–1216, New York, NY, USA, 2008. ACM.

[84] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5, 2007.

[85] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, 2008. ACM.

[86] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, 2008. ACM.

[87] S. Levy. How many friends is too many? *Newsweek*, May 2008.

[88] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. In *Proceedings of National Academy of Sciences*, pages 623–628, 2005.

[89] A. Markwick. I'm more than just a friendster profile: Identity, authenticity, and power in social networking services. 2005.

[90] J. P. Mazer, R. E. Murphy, and C. J. Simonds. I'll see you on "facebook:" the effects of computer-mediated teacher self-disclosure on student motivation, affective learning, and classroom climate. *Communication Education*, 56(1):1–17, 2007.

[91] A. M. McQueen. Trends: Facebook – too much information! *Sun Media*, August 2007.

[92] S. Milgram. The small world problem. *Psychology Today*, (2):60–67, 1967.

[93] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA, 2007. ACM.

[94] C. A. Murnan. Expanding communication mechanisms: they're not just e-mailing anymore. In *SIGUCCS '06: Proceedings of the 34th annual ACM SIGUCCS conference on User services*, pages 267–272, New York, NY, USA, 2006.

[95] V. Murphy. You're not my friendster. *FORBES*, 172(12):59, 2003.

[96] M. E. J. Newman. Assortative mixing in networks. *Physics Review*, (89), 2002.

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

[98] M. E. J. Newman and M. Girvan. Finding and evaluting community structure in newtworks. *Physics Review*, (69), 2004.

[99] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physics Review*, (68), 2003.

[100] R. Nyland and C. Near. Jesus is my friend: Religiosity as a mediating factor in in-
       ternet social networking use. In *Paper presented at AEJMC Midwinter Conference*,
       Reno, NV, 2007.

[101] I. O'Murchu, J. G. Breslin, and S. Decker. Online social and business networking
       communities. Technical report, DERI Technical Report, August 2004.

[102] R. Pastor-Satorras and A. Vespignai. Immunisation of complex networks. *The
       American Physical Society*, 2002.

[103] S. Patil and J. Lai. Who gets to know what when: configuring privacy prefer-
       ences in an awareness application. In *ACM Conference on Human Factors and
       Computing Systems*, Portland, OR, 2005.

[104] D. Pattishall. Friendster scales-out with mysql network. *White Paper*, 2005.

[105] D. Pattishall and C. Lunt. Friendster scales-out with mysql network. *Database
       and network Journal*, (36):15, 2006.

[106] D. Reed, M. L. Maitre, B. Barnhill, O. Davis, and F. Labalme. The social web:
       Creating an open social network with xdi. *Planetwork Journal*, 2005.

[107] R. H. Reido. *Architects of the Web: 1000 days that built the future of business*.
       New York: John Wiley and Sons Inc., 1997.

[108] P. Resnick. Beyond bowling together: Sociotechnical capital. *HCI in the New
       Millennium*, pages 247–272, 2001.

[109] D. Rosenblum. What anyone can know: The privacy risks of social networking
       sites. *Security and Privacy, IEEE*, 5(3):40–49, 2007.

[110] J. Scott. *Social Network Analysis: A Handbook*. Sage Publications, London, 2nd
       ed., 2000.

[111] O. Simsek and D. Jensen. Decentralized search in networks using homophily and
       degree disparity. Proc. 19th International Joint Conference on Artificial Intelli-
       gence, 2005.

[112] J. Snyder, D. Carpenter, and G. J. Slauson. Myspace.com - a social networking
       site and social contract theory. Proceedings of ISECON, 2006.

[113] E. Spertus, M. Sahami, and O. Buyukkokten. Evaluating similarity measures:
       a large-scale study in the orkut social network. In *KDD '05: Proceeding of the
       eleventh ACM SIGKDD international conference on Knowledge discovery in data
       mining*, pages 678–684, New York, NY, USA, 2005.

[114] S. Strogatz. Exploring complex networks. *Nature*, (410):268, 2001.

[115] F. Stutzman. An evaluation of identity-sharing behavior in social network communities. *Journal of the International Digital Media and Arts Association*, (3): 10–18, 2006.

[116] K. Swinth, S. Farnham, and J. Davis. Sharing personal information in online community member profiles. *Internal paper*, 2002.

[117] E. L. Thorndike. Intelligence and its use. *Harper's Magazine*, 140:227–235, 1920.

[118] S. T. Tong, D. H. Brandon, L. Lwell, and J. B. Walther. Too much of a good thing? the relationship between number of friends and interpersonal impressions on facebook. *Journal of Computer-Mediated Communication*, 13(3):531–549+, 2008.

[119] P. M. Valkenburg, J. Peter, and A. P. Schouten. Friend networking sites and their relationship to adolescentsẃell-being and social self-esteem. *CYBERPSYCHOLOGY & BEHAVIOR*, 9(5):584–590, 2006.

[120] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. New York, Cambridge University Press, 1994.

[121] D. J. Watts. *Six degrees: the science of a connected age*. Norton, New York, 2003.

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

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

[124] B. Wegener. Job mobility and social ties: Social resources, prior job, and status attainment. *American Sociological Review*, 56(1):60–71, 1991.

[125] D. Williams. On and off the 'net: Scales for social capital in an online era. *Journal of Computer-Mediated Communication*, 11:247–272, 2006.

[126] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *EuroSys '09: Proceedings of the fourth ACM european conference on Computer systems*, pages 205–218, New York, NY, USA, 2009. ACM.

[127] L. Wu, C. Y. Lin, S. Aral, and E. Brynjolfsson. Value of social network – a large-scale analysis on network structure impact to financial revenues of information technology consultants. In *Winter Information Systems Conference*, 2009.

[128] A. Zinman and J. Donath. Is britney spears spam? *Paper presented at the Fourth Conference on Email and Anti-Spam*, 2007.