UNIVERSITY OF SOUTHAMPTON
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

A mini-thesis submitted for transfer from MPhil to PhD

Supervisor: Prof. Nigel Shadbolt
Examiner: Dr Nicholas Gibbins

Supporting Meaningful Social Networks

by Yongjian Huang

October 1, 2007

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

A mini-thesis submitted for transfer from MPhil to PhD

by Yongjian Huang

Recent years have seen rapid growth of social networking sites such as Friendster, MySpace and Facebook. As they grow, there are many issues emerging from people's online interactions. The problems we examine include online personas, acquaintance relationships, fakesters, contact searching, trust and privacy. The thesis proposes a model based on the Barabasi-Albert network to predict the growth and evolution of social networking sites. We find that online social networks fail to maintain the scale-free topology of power law degree distribution when the average number of friends of 44.78% of the users passes Dunbar's number, which is the maximum number of individuals with whom any one person can maintain stable relationships. Thus, we present RealSpace, a social networking site based on a dynamic social network model. Our goal is to establish the online community by capturing the connections in the real world. An important feature of the dynamic model is to eliminate socialising footprints automatically. Also, we discuss the problems of search in social networks. We compare two search paradigms: exhaustive search and decentralised search. A ranking mechanism based on *social connectivity* is designed to improve the ranking quality for people searching. A novel algorithm for decentralised search based on social distance with transferrable social tables has been proposed to overcome the weakness of centralised exhaustive search. A prototype is presented to show the implementation of the system. Finally, we look at the problems on how to improve the network model and decentralised search performance.

Contents

Chapter 1

Introduction

People share information and communicate. Tools for information sharing and communication will be developed and deployed quickly if they are more efficient, cost-effective and user-friendly. Email was developed before the Internet and quickly became popular because it was more efficient and cheaper than traditional letters. People rapidly embraced the new technology, moving from telephone communication as the primary means, to email communication[28]. Instant messaging applications began to appear in the 1970s and modern GUI-based instant messaging clients such as ICQ were not developed until mid 1990s. Yet it quickly took off because it offered real-time communication and facilitated efficient collaborations. Web-based blogs emerged as a new way of communication in the early 2000 and attracted much attention. VoIP such as Skype based on peer-to-peer technology provided high quality yet low cost of voice communication and took the world by storm.

Recent years have seen the dramatic growth of social networking sites such as Friendster, MySpace and Facebook. Figure 1.1 shows how social networking sites grow rapidly in the last five years. Friendster was founded in 2002 and gained huge popularity in 2004. MySpace was founded in 2003 and is now the third most popular site in the US behind Yahoo and Google. Facebook was founded in 2004 and has the largest number of registered users in the college. There are three advantages of social networking services over previous information sharing and communication tools. First, they are built on top of the Web platform, which is very ubiquitous and easily accessible. This makes SNS reached by more people. It is very convenient for users to discover new friends and trace their credibility in the network based on the Web. Second, they allow asynchronous communication and information sharing with multiple parties instantaneously. One publishes his information once and all of his friends can receive it. Third, Web-based social networking sites are capable of accommodating and integrating other communication technologies. Users of social networks can normally communicate with each other by sending messages inside the network. This can be compared with traditional email system. Many social networks also provide blogging. Recently, MySpace introduced

MySpaceIM, an instant messenger that uses one's MySpace account as a screen name[1]. It is even possible for resource-heavy applications such as VoIP services to be deployed in social networks when browsers implement peer-to-peer distribution protocol[2]. These edges of SNS help people to share information and communicate more efficiently than ever before. They give people a more powerful tool to discover new friends, strengthen and maintain existing relationships. Hence, social networking sites enjoy a rapid growth and go from strength to strength.

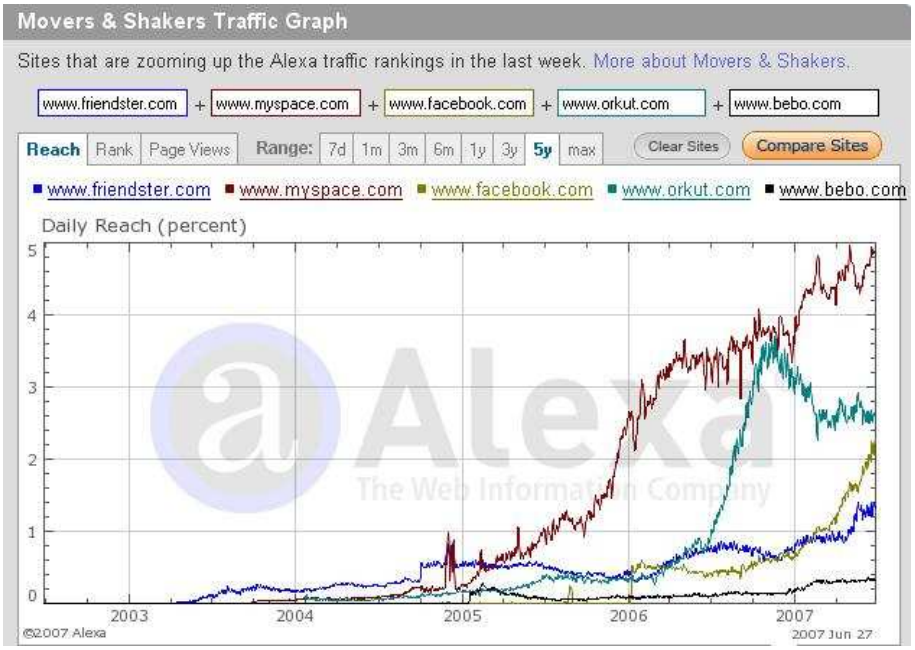

FIGURE 1.1: The Traffic of Friendster(2002), MySpace(2003), Facebook(2004), Orkut(2004) and Bebo(2005) from 2002 to 2007

1.1 History of Social Networking Services

Social networking sites attract much attention but the idea is not new. The history of social networking sites can be dated back to 1994, a year after the historical Web browser Mosaic was released. The development of social networking sites reflect people's efforts to connect with each other through the Web.

1.1.1 1994-2000: Circle of Contacts

In 1993, the Mosaic browser, which was GUI-based and worked on Microsoft Windows operating system, was released. Unlike previous browser which mainly worked with texts, Mosaic could handle a mixture of texts and graphics. The Web browser transformed the appeal of the Web from academic uses in the technical area to mass-market

[1]MySpace: http://www.myspace.com/myspaceim
[2]MozTorren: http://moztorrent.mozdev.org/

appeal[31]. Since then, people had started to explore the new platform to publish and access information. The websites in the early years included *Amazon*, *Yahoo!*, *eBay* and *Match.com*. *Match.com* was an online dating site. The website maintained the contacts and profiles of the members which others could search and keep in touch with. The method was known as *Circle of Contacts* and was used by most of the social networking sites at that time. The problem of the method was that it did not allow people to add contacts and communicate on the websites directly. Users had to communicate with other users either by email or offline. The limitation was mainly because in the early years of Web development, the Web standards, such as HTML, Javascript and Scripting languages were very primitive, many of which had not been developed yet. Still, many online communities were created, including *Student.com* and *Classmates.com*, which facilitated socialising for college students and alumni, in a way that similar to today's Facebook.

Inspired by the social theory of *six degree of separation*, *sixdegrees.com* was created in 1997. It was the first social networking site which was built based on social network theories. The website pioneered the concept of *small world phenomenon* on the Web and encouraged many people to use online social networking services. Despite the great enthusiasm, the site was still using the inefficient method of *Circle of Contacts*, it was not able to offer more efficient ways for socialising. *Sixdegrees.com* stopped to function in the recent years. The enthusiasm was effectively killed by the ineffective technology.

1.1.2 2000-present: Circle of Friends

By the year 2000, the technologies had been mature enough to break the bottleneck. Broadband Internet access was available and widely used; desktop PCs with GUI-based user friendly operating system were commonplace; web browsers were very established; HTML 4.01 was made available and RSS 0.9 published. The advent of the technologies incubated a new way of online socialising: *Circle of Friends*. People could send a friend request and add friends on the Websites. The improvement enabled people to acquire new contacts and communicate much efficiently. The method is widely used as Friendster gained huge success and attracted massive press coverage[7]. Gradually, the social networking sites not only enabled friend request and invitation but also made direct communication through the sites available. The method became mainstream in the following years, influencing the development of subsequent social networking sites, such as MySpace and Facebook.

The major players in the industry have also embraced and adopted SNS due to its huge popularity and commercial success. Google launched Orkut in 2004[34]. Yahoo! 360 was developed in 2005. Microsoft recently renewed its social networking platform, Windows Live Spaces.

1.2 The Social Networking Platform

Social networking services move people's relationships from real world to cyberspace. They provide a new platform for people to share information and communicate. At the end of May 2007, having already opened some APIs to third party developers, Facebook announced that the company was going to develop the social networking site as a platform for programmers to develop applications, in the same sense as developing applications on the computer platform and Web platform[3]. Figure 1.2 compares the three platforms.

The idea of evolving social networking services to social networking platform advances the way people share information and communicate to a higher level. The social networking services are now not only a platform for people to socialising but also platforms for people to develop applications to publish and spread information. In fact, contemporary social networking services have already provided many tools for information sharing such as blogs, group discussion, video and music sharing. They offer even better choice for people to spread knowledge and information than the Web itself, thanks to the power of viral marketing and advertising[20]. Numerous studies have shown that one of the most effective channels for dissemination of information and expertise within an organisation is its informal network of collaborators, colleagues and friends[21]. Some sources indicated that photo sharing on Facebook was more popular than Flickr[4] and the South Korea-based social networking site, Cyworld, which offers blogging, music and video sharing, claimed to have more traffic than the highly touted YouTube[5]. The power of social networks in spreading information explains why singers promote their music albums through MySpace.

Soial networking sites, though in their infancy, may provide a completely new mechanism for information technology. IBM PC came to the market in 1981 and quickly became a huge commercial success. The word processors *WordStar* and the database *dBase* were introduced to DOS platform from CP/M platform. The spreadsheet program Lotus 1-2-3 was written for the platform. These software provided cost-effective tools for people to publish information in the right form. Before the World Wide Web was invented, BBS, Usenet and Gopher used to be information sharing systems on the Internet. Since the Web was developed in 1991, those information sharing tools were quickly surpassed by the Web because it offers more flexible and faster ways to publish and access information. The killer app Mosaic browser brought the Web to generate public, making information accessible in the right form at the right time. With the popularity of online services such as Google, Wikipedia, Youtube and MySpace, Web 2.0 is currently in the lead to move applications from computer platform to Web platform, allowing services and software to deliver instantaneously. Built upon the Web platform, social networking sites

[3]Facebook Platform Launches: http://developers.facebook.com/news.php?blog=1&story=21
[4]Facebook Blog: http://blog.facebook.com/blog.php?post=2406207130
[5]Cyworld News: http://www.usnews.com/usnews/biztech/articles/061109/9webstars.cyworld.htm

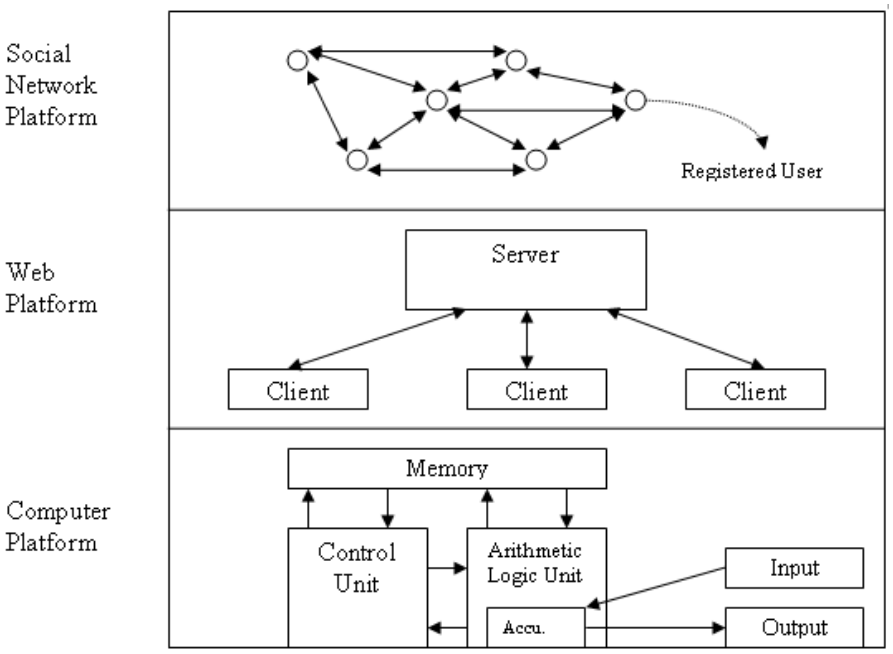

FIGURE 1.2: The Social Networking Platform

are able to facilitate more efficient information sharing and communication, enabling information and knowledge to spread in the right form at the right time to the right people.

The remaining chapters are arranged as follows: Chapter 2 presents a literature review of research on social networks. These previous studies provide a theoretical framework for our research on social networking sites. In chapter 3, we will discuss the issues and problems in social networking services. A model is proposed to predict the growth and evolution of social networking sites. Chapter 4 illustrates the design of RealSpace system. It includes the dynamic structure of the social networks and two search algorithms: exhaustive contact search and decentralised search. In chapter 5, we give the architecture of the system and show the work we have completed. Finally, in chapter 6, we will look at the problems in future research.

Chapter 2

Reviews of Social Networks

Recent advancement in online social networks may be a new phenomenon but research on social network have been studied extensively in the last decades. Social network is often referred to a type of complex networks, which have certain non-trivial topological features that are not present in simple networks. One of these characters is small-world effect. It also A famous experiment was carried out by Stanley Milgram in the 1960s, in which letters passed from person to person were able to reach a designated target individual in only about six steps[27]. The result is one of the first direct evidences of small-world effect. The experiment also incite huge interests and enthusiasm in social network. In the past few years, the progress of information technology led to the emergence of large databases on the topology of various real social networks. Computing powers allowed researchers to investigate networks containing millions of nodes, exploring questions that could not be addressed before. Thus, many new concepts and measures have been proposed and investigated in depth in the past few years. So far, the study on various data largely confirms that the networks have three robust measures of topology[2][29]: small average path lengths 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 properties. 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[35]. 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[38]. It is known as transitivity in sociology[36]. Transitivity, which is derived from *balance theory*, is proposed as a fundamental social law[36]. For power law distribution, Albert, Jeong and Barabasi suggested that scale-free networks are resistant to random failures because a few hubs dominate their topology[3]. The existence and persistence of these interesting characters in social network as well as other complex networks inspired researchers to construct new mathematical models for network study.

2.1 The Erdos-Renyi Model

In their classic article on random graphs, Erdos and Renyi proposed a simple model of a network. Take some number of n nodes and connect each pair with probability p. This defines $G_{n,p}$ in the ER model[13]. Figure 2.1 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 indicates the small-world effect. However, the model fails to describe other significant features such as clustering and degree distribution.

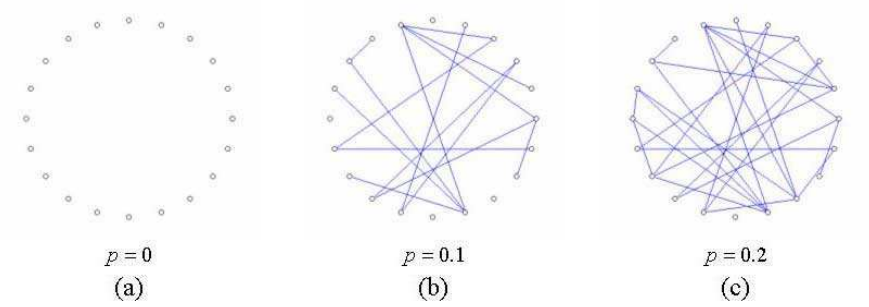

FIGURE 2.1: Illustration of the graph evolution process for the ER model

2.2 The Watts-Strogatz Model

The real-world social networks have a small-world character like random graphs, but they also have exceptionally large clustering coefficients, which was not 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 2.2): Starting from a ring lattice with n vertices and k edges per vertex, each edge is rewired at random with probability p[40]. Watts et al. found that $L\ n/2k \geq 1$ and $C\ 3/4$ 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 and it concludes that the WS network is suitable in explaining such character in real network.

The model has been studied widely since the details have been published. 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 the it has not been able to capture the power law distribution as demonstrated in real social networks.

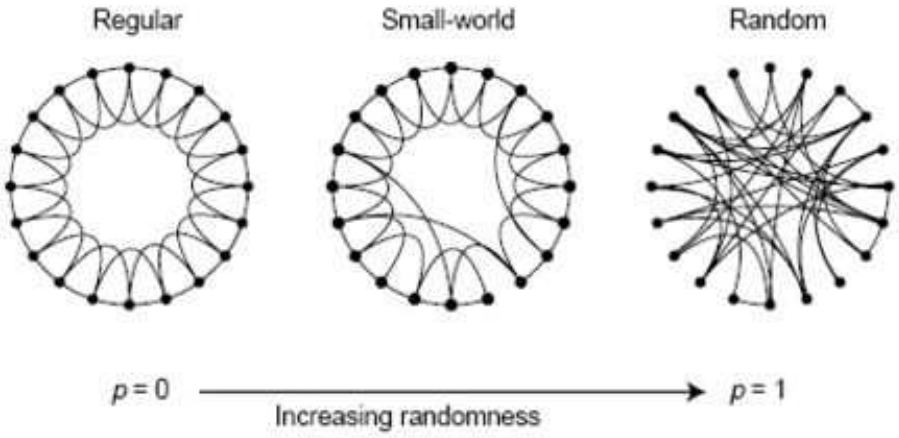

FIGURE 2.2: The random rewiring procedure of the WS model which interpolates between a regular ring lattice and a random network

2.3 The Barabasi-Albert Model

The previous two models take observed properties of real-world networks and attempt to create networks that incorporate those properties. These models do not help understand the origin of social networks and how they generate those properties as they grow. Barabasi and Albert proposed a model that tried to address these problems. There are two important hypothesis with the model[5]:

(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 $\frac{1}{2}\sum_k kp_k$. 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 present in the system.

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

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

Using master-equation approach they show:

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

It has been pointed out that the concept of *preferential attachment*, is largely influenced by the notion of *cumulative advantage* in Price's model[29]. In the limit of large k it gives a power law degree distribution $p_k \sim k^{-\alpha}$, with the $\alpha = 3$. Figure 2.3 shows the degree distribution for the model. While the BA model captures the power law tail of

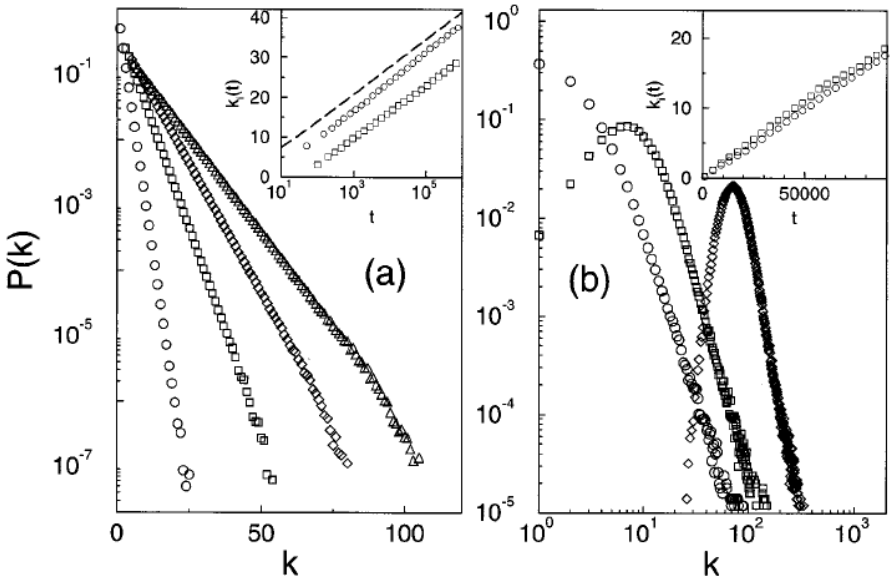

FIGURE 2.3: Two Simulations for Degree Distribution for BA model

the degree distribution, it has other properties that may or may not agree with empirical results on 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 a random graph. The clustering coefficient decreases with the network size, following approximately a power law $C \sim N^{-0.75}$. Though greater than random graphs, it is not independent of network size, which is not true for real-world social networks.

Two limiting cases have been conceived to test the two hypothesis 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, 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 Search on Networks

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.

First, we clarify the concept of search on networks. Suppose some resource of interest stored at the nodes of a social network, such as expertise of engineers and information held by individuals. One would like to find out rapidly where on the network a particular item of interest can be found. Milgram's experiment, as mentioned above, does not just shows us that small-world effects exists in social network, but also tell us ordinary people, using only local information, are able to construct short paths to find the target. These observations led to the influence of the network topology on the search behaviour.

2.4.1 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[23]. The model employs a two-dimensional lattice (with size $n \times n$) as basic structure. Notice that it was NOT a ring 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 know his 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 2.4 shows the graph of the lattice.

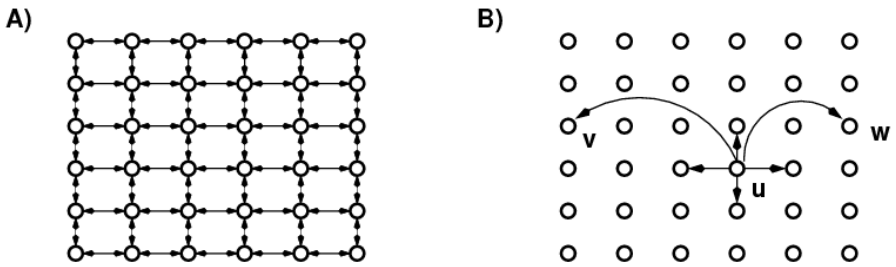

FIGURE 2.4: (A) A two-dimensional grid network with n=6, p=1, and q=0; (B) p=1 and q=2, v and w are the two long-range contacts

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 his 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[24]. Kleinberg's proof reveals an important feature of search on social networks: the existence of short paths lies not on the sophistication of search algorithm 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.

2.4.2 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 target individuals[22]. They pointed out that the top choices for selecting a friend is location and occupation. Watts *et al.* proposed a model for a social network that is based on social grouping[39]. 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 2.5.
(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 true distance; network paths, which generate 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

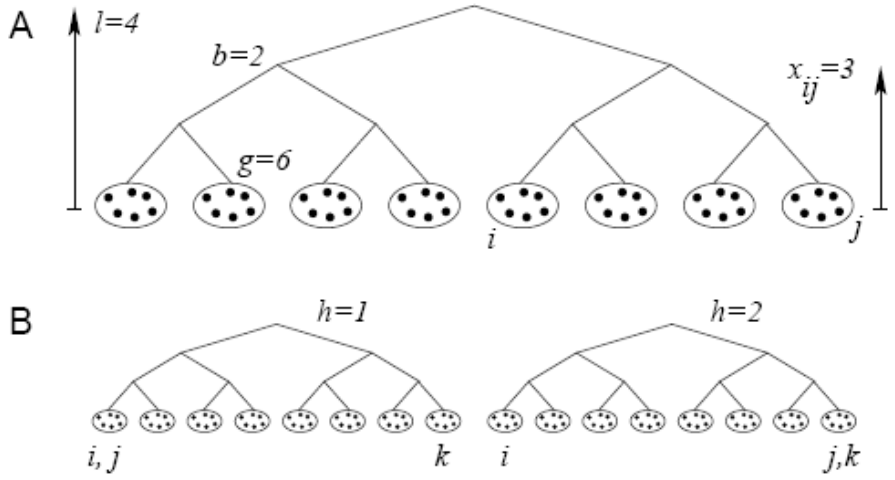

FIGURE 2.5: The Hierarchical "Social Distance" Tree Model

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[24]. Based on the result of computer simulation, Simsek and Jensen[33] suggested that a heuristic decentralised algorithm taking both social distance and node degree information can perform more efficiently than using only one of these factors.

Chapter 3

Growth Constraint of Social Networking Sites

As we argued in the first chapter, social networking site offers a technologically advanced tool for people to share information and communicate. However, the relationships between persons are much more complex and dynamic than the relationships between servers and clients or CPU and memory. Social networks can be perceived as an open multi-agent system (MAS) of most intelligent agents and of most sophisticated relationships. As the network grows, there are many issues emerging from people's online interactions. These issues further compound by the fact that companies who own the social networking sites have a tendency to design and maintain their social networking services based on business interests. We discuss in details the problems confronting many social networkings sites. Based on these facts, we propose a model to explain the growth and evolutions process of online social networks. The model predicts that there is an inevitable growth constraint in most social networking sites.

3.1 Issues in Six Degrees

3.1.1 Online Personas

Most social networking sites allow people to present themselves through a profile. A typical profile includes name, gender, location, interests, education background and profession. Some may also display information about their social network, relationship status, contact methods, etc. Figure 3.1 shows a profile from Facebook.

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 efforts to make the profiles

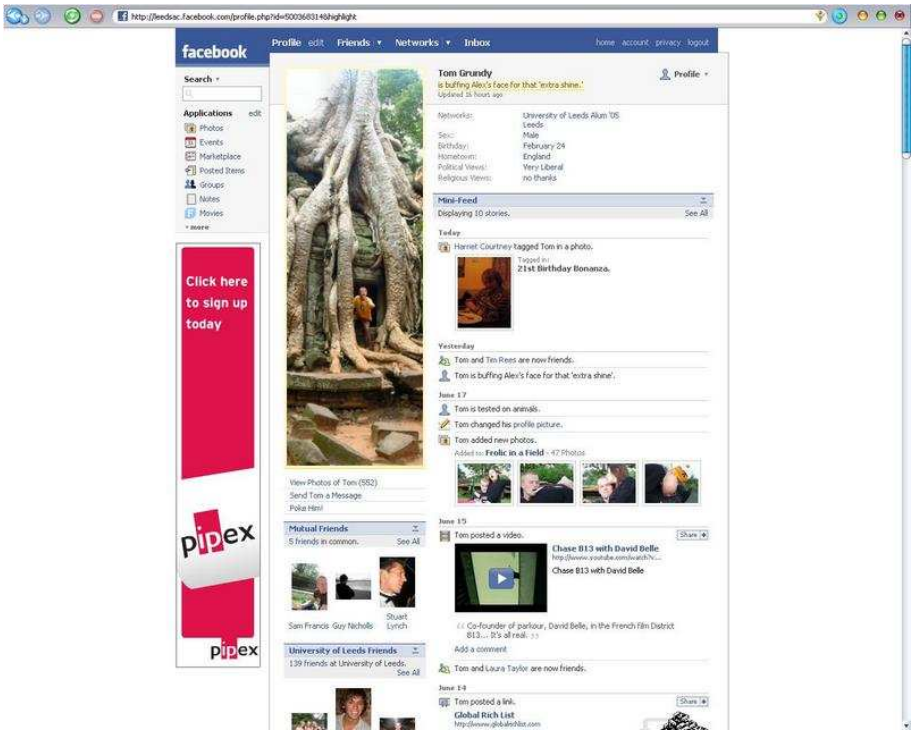

FIGURE 3.1: Facebook Profile

represent themselves as accurately as possible. For example, a research suggested that users in the Facebook "reported high confidence that their Facebook portrayals described them accurately and that those portrayals were positive"[26]. But it is not unusual for people to take photos of celebrities and put them in their profiles.

In our daily life, we usually present ourselves differently with different audiences and we will try to behave appropriately in different situation and context[15]. When social networking sites move people's relationships to cyberspace, users also bring their various social masks online. However, most sites do not yet provide tools for managing multiple profiles and masks. The communication goes well when only a specific group of people using the sites but it will cause problems when more users and audiences from different backgrounds join in. In Facebook, for example, users view their audience as peer group members, but not faculty, administration within the campus, or outsiders. Thus, they behave in a way similar to what they do in student community. 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 in Friendster that most users fear the presence of two people: boss and mother[31]. Research also suggested that teachers fear the presence of their students. Social networking sites usually address this problem by giving users control of their profiles by defining 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 a 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 networking sites, on the contrary, are much less context sensitive. They are eager to attract more users but fail to provide tools to accommodate multiple online personas. The sites will become an increasingly embarrassing socialising place when relationships and interactions become more diversified.

3.1.2 Acquaintances as Friends

In the social network, it is very easy to 'make friends'. *Friends* is the term applied to members of a social network who list on someone else's page. To add someone as a friend simply takes a single click. In such a world where friendships are mediated through a digital interface, it is not difficult for any new comers to any social network to acquire a large amount of 'friends' in a short time. Figure 3.2 shows a user and her friends on Orkut. The user has 382 friends, most of whom have more than 200 friends. One even has over 700 friends.

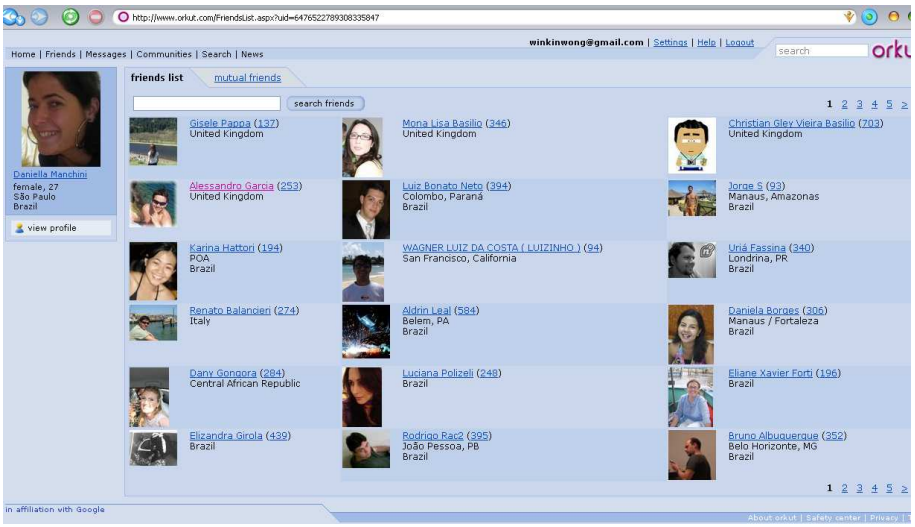

FIGURE 3.2: 'Friendster' Friend

The easy aggregation of friends is the benefit from the new method *Circle of Friends*. However, the drawback is, with so many friends and relationships, there is no way to determine what metric is used or what the role of weight of the relationship is[7]. Danah Boyd observed in her research 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. This often means that people are indicated as friends even though the user does not particularly know or trust the person. In some cases, it is necessary to publicly be-Friend someone simply for political reasons. Sometimes, people connect broadly so that they may see a larger percentage of the

network". Some teens will accept total strangers as friends in an attempt to boost the total number of friends noted on their page, and so appear popular[4]. Because of the weakness in the system, some users on *Friends.com* use the term 'friendster' to signify the casual acquaintance[7].

Research in sociology indicates that there is the maximum number of individuals with whom any one person can maintain stable relationships. This is commonly known as Dunbar's number, which is about 150[11]. Based on the theory we can conclude that presence of loose acquaintances is inevitable in one's network where the number of friends is over 150. On the other hand, the *Weak Tie Hypothesis* suggests that casual acquaintances are likely to introduce new ideas and bridge different networking groups and cliques[16]. Therefore, users of social networks will probably try to maintain in their contact lists as many acquaintances as possible, even though they do not interact with them. They are not willing to remove the acquaintance contacts unless there is an explosive end to the relationship[7].

Owners of social networks see the benefits and drawback of acquaintances. Some try to set a limit to the number of friends one can have while others provide tools and facilities for users to categorise and manage the relationships. In Orkut, for instance, friends can be organised as an acquaintance, friend, good friend, best friends, or someone you have not met[1]. In Facebook, relationships are classified in details. The spectrum of relationship status in Facebook is listed as follows:

- Lived together;
- Worked together;
- From an organisation or team;
- Took a course together;
- From a summer / study abroad program;
- Went to school together;
- Traveled together;
- In my family;
- Through a friend;
- Through Facebook;
- Met randomly;
- We hooked up;
- We dated;
- I don't even know this person.
- From an organisation or team;

The assumption of fine-grained categorisation is that users are willing to spend time on categorising their friends and be accurate and honest about such editing. Unfortunately, the practice is usually ignored by the users. We observe that users either ignore such functions as they are too complicated, or define in the relationship without much attention to the accuracy. When users do specify the relationship, the description may

[1]Contact Management in Orkut: http://help.orkut.com/support/bin/answer.py?answer=11765&topic=10315

not reflect the real relationship. For instance, 'Went to school together' relationship is supposed to be closer than 'Through a friend'. However, this may be false if friends in the former case do not contact with each other after graduation but friends in the later case keep close contact on a regular base. Worse still, as users' network evolves over time, the relationship will change correspondingly. But users are not keen to update the relationship in the system even though they know the change.

The manually editing mechanism is subject to abuse. As analysed above, people might be-Friend with someone simply for political reasons. If they see the value of putting their friends in certain category, they may try to manipulate the relationships. For example, in Orkut, they can deliberately put their friends in 'best friends' category in order to show them off. Hence, acquaintances still coexist with friends in most users' contact lists. It is very difficult for current policy and techniques to solve the problem.

3.1.3 Fakester Dilemma

Similar to acquaintance problem, the term *fakesters* emerged from early social networking site, Friendster. Fakesters are fake personas created by users for different purpose. Figure 3.3 shows a fakester, Tony Blair and his fakester friends on MySpace.

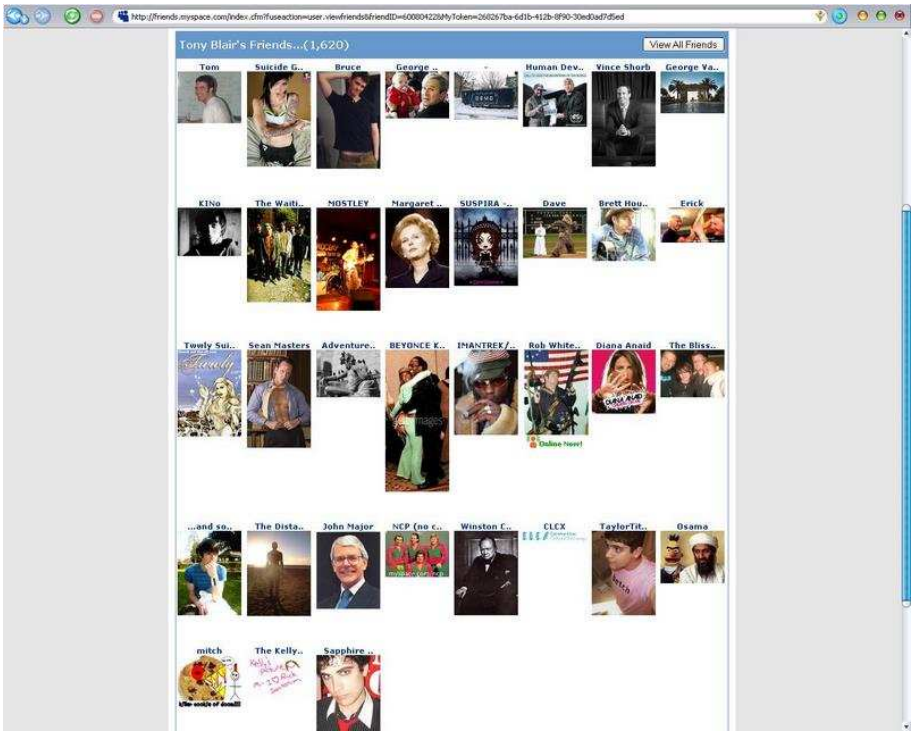

FIGURE 3.3: Fakesters on Myspace: Tony Blair's Friends

Research on Friendster revealed three categories of fakesters[7]:

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.

Fake identities become increasingly common on Facebook, which has been open to public register[17].

TABLE 3.1: Fake Profiles on Facebook

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

The fakester phenomena reflects the dynamics of the users. SNS users are extremely active in creating fakesters. However, owners of the social networking sites dislike fakesters. They argued that these fake profiles will collapse the network, devaluing the meaning of connections between people. Some companies, such as Friendster, has attempted to eliminate all of these fake users by removing them from their sites and servers. This, however, created tension between the company and users[7]. The company saw massive rebellions from the users. In fact, many users love fakesters and the fakester culture have spreaded to many other social networking sites.

3.1.4 Contact Searching

Search is most studied on the Web and many search and ranking algorithms have been proposed and developed. The query interface of Web search engine is very simple – one queries by inputing keywords. In contrast, social networking sites provide more sophisticated query mechanism. In social networking sites, members can be searched by specifying an array of conditions. Users can issue queries by searching name, gender, location, interests, education background, career information. Figure 3.4 shows the advanced search interface of Facebook.

However, when it comes to ranking the search results, social networking sites are lack of effective solution. Many sites rank the people in alphabetical order of their surnames. Some sites try to filter the results by certain criteria, such as friends of friends (Orkut, see Figure 3.5), relationships and friends' recommendation. These approaches are useful but they have two major disadvantages: first, as the number of members in the community increases, the search could lead to a list with hundreds and thousands results which are not desired. Second, sorting strategy based on human ratings is inefficient and subject to manipulation. Therefore a more robust and objective search engines are needed to handle contact searching efficiently.

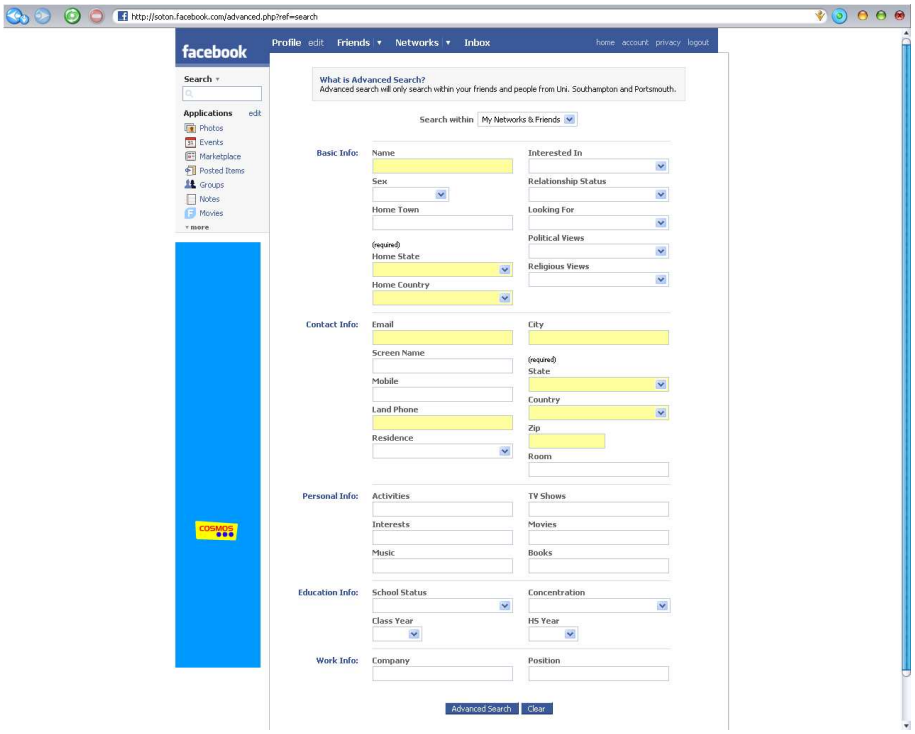

FIGURE 3.4: Advanced Search on Facebook

3.1.5 Trust and Reputation

Given the dynamics of the social network, the complexity of people's real relationships, how can we determine what person we can trust in the cyberspace? In our own network, we can distinguish acquaintances and friends and we trust our friends more than acquaintances. But how about those acquaintances? After all, they are weak ties and may as well be important to us. For the rest of the members in the social network, how should we make the decision on how trustworthy they are and how can we justify our judgements?

Common sense tells us that friends of friends are the people we can trust, as we can trace their credibility. In fact, *Friendster.com* is founded based on this idea[7]. The problem with this approach is that, on the online social networking sites, users do not know immediately if the friends of their friends are close friends or loose acquaintances. As pointed out previously, there is no effective metric to distinguish acquaintances from friends beyond one's own network. Thus, users have to ask their friends explicitly to elicit the relationships in their friends' networks. This strategy may be reliable but usually inefficient and time-consuming.

Some social networks allow users to rate their friends, in the same way as online auction websites and recommendation systems do. In fact, reputation systems are used extensively by auction sites to prevent fraud. Research on reputation systems, however, suggested that such systems face many challenges which include the difficulty to elicit

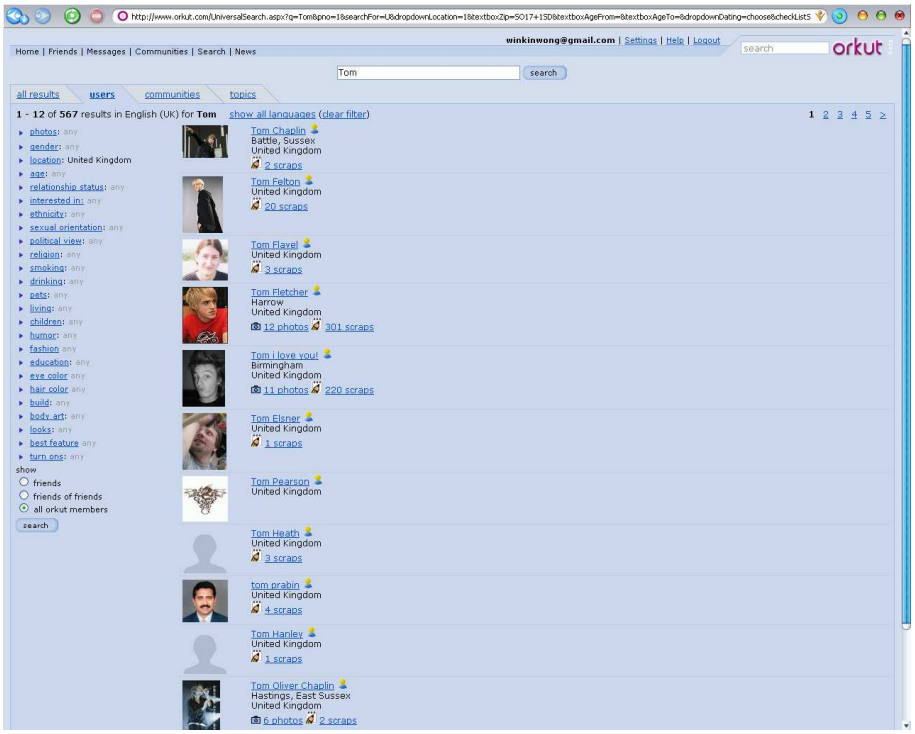

FIGURE 3.5: Filters on Orkut

honest feedback and to show faithful representation of users' reputations[32]. This is particularly true in social networks, where people share information and interact with others for different purposes. Thus, it is contrast to auction sites where users only rate the sellers based on the any single transaction.

In most cases, we may simply look at the profiles of the users and make the judgement based on the information about them. This is similar to the first approach, which might be effective but not efficient. As the network grows, it cost more and more time for a user to examine all the information about the target individual. In summary, in the social networking services, there are currently no objective efficient metrics for indicating the reputation of the members and the degree of trust.

3.1.6 Privacy Concerns

People share significant amount of information on social networks. Due to the dynamics of social networks, as well as the issues of online personas, acquaintances, fakesters and trust, online privacy is always a pressing concern. Privacy can be defined as "the freedom from undesirable intrusions and the avoidance of publicity". There is always a tradeoff between personal privacy and our desire to share information and communicate. In certain occasions we want information about ourselves to be known only by a small circle of close friends, but not by strangers. In other instances, we are willing to reveal personal information to anonymous strangers, but not to those who know us very well.

Over the past decade, surveys have consistently reported that people express high levels of concern about online privacy. 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[14].

Despite these concerns and the common sense caution about sharing personal information, users are still willing to share a large amount of their personal data in the social networks. This seems to be a paradox. A number of factors have been identified to be likely to drive information revelation in online social networks[17].

1. Perceived benefits of selectively revealing data to strangers may appear larger than the perceived costs of possible privacy invasions;
2. Peer pressure and herding behaviour;
3. Relaxed attitudes towards (or lack of interest in) personal privacy;
4. Incomplete information (about the possible privacy implications of information revelation);
5. Faith in the networking service or trust in its members (In Facebook, vast majority of users are college students with university email accounts and therefore the network is perceived as safe and trustworthy);
6. Myopic evaluation of privacy risks.

In practice, very few people have had anything seriously harmful happen to them while online[14]. Thus, social networking sites, which facilitate the exchange of personal information, are booming in popularity.

However, something that does not happen doesn't suggest that it will not happen. The exposure of personal information is easily subject to abuse by malicious elements in the society. There are at least three types of misuse of personal information[30]: first, it is possible that social networking sites will exploit user profile information to mine data for targeting specific advertisements. Personal information related to consuming behaviour is particularly in great interest of advertising and marketing industry. Second, the low entry barrier to social networking 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 a more trustworthy place due to the presence of their real-world friends, more information about personal identity can be found and misused. Third, some sites state in their privacy policy that they may provide personal information to a third party in order to facilitate or outsource aspects of their services, though users usually ignore the terms and conditions when they sign up with a site. Information sharing with third parties might provide better services to users, as many site claim, but it equally incurs risk of privacy leaking.

It has been argued that it is vitally important to grant the users full control of their own information. In practice, most social networking sites provide privacy settings. Some

sites, such as Facebook, have sophisticated privacy options that let the users choose who can have access to their personal information. The problem is, the default settings of many social networking sites are likely to reveal as much information as possible, since owners of the social networks want to attract more people to register. On the other hand, users do not usually change the settings. This is because they do not bother to change it or they do not know how to change it. Lack of fine-grained privacy settings may also be a problem.

3.2 Accumulative Network Model

Almost all the social networking sites will retain the relationships that a user has made since he registered. Users themselves tend not to remove the historical relationship links or acquaintance links even when the number of their links go far beyond the one they can keep contact with. Thus, the topology of the online social network is a combination of people's real connections and socialising footprints, which refers to the friends in socialising history and loose acquaintances. We may call such networks accumulative networks. They are different from the real-world social network in that socialising footprints have been retained and kept accumulating. We ask, does this type of network have the same topological features, such as small-world effect, large clustering and power-law distribution, as in the real social network? To answer the question, we propose a model to simulate the growth and evolution of the accumulative network.

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

(a) In BA model, the exponent α=3, but in real network, the number is between 2 and 3. We use 2.3, which is the measure for movie 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 people are capable to keep regular contact with at 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 bigger than or equal to 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 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.

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

$$p_k = 2m(m+1)k^{-2.3} = 45300k^{-2.3} \tag{3.1}$$

Figure 3.6 shows the graph of Eq 3.1.

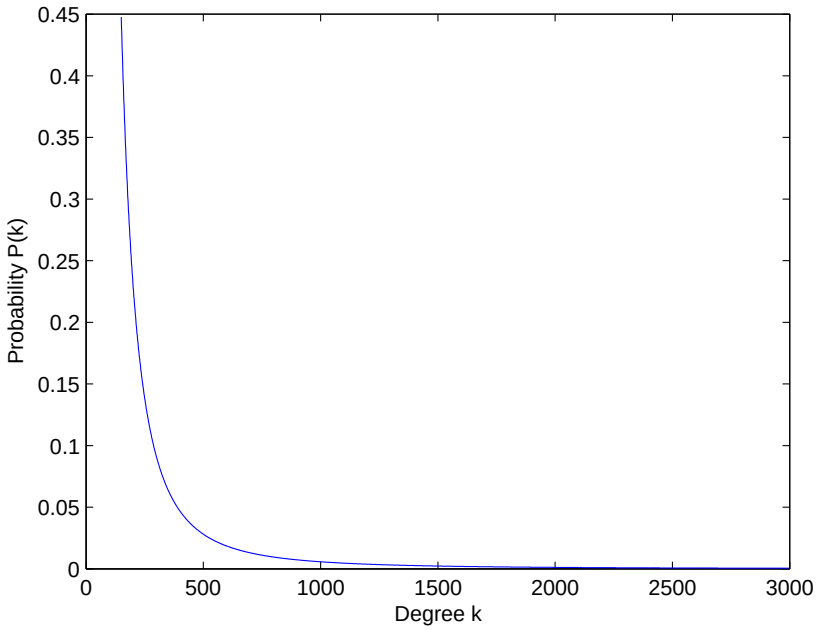

FIGURE 3.6: Degree Distribution in BA model with m=150, α=-2.3 but no edge rewiring

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 can make friends over 150. 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, neighbourhood that you know and people whom you have worked or studied with for some time. So far, empirical data shows none of the social networking 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 networking 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 accumulative networks, the footprints will not disappear automatically, which is contrast to the real social networks where old relationships will be forgot gradually when people do

not keep regular contact with each other. We discuss two scenarios of the consequences for the development of social networking sites:

Scenario 1: as the number of friends goes beyond 150 and continues to grow, it is not uncommon to find people who have thousands of friends. In the real world, nevertheless, people with many contacts are usually restricted to the rich, politicians, celebrities and leaders. Common people would like to make friends with these high-profile figures, but find it very difficult to do so. However, in social networking sites, the notion of high degree simply does not imply the high social status of the individual. This will destroy the factor of *preferential attachment* as described in BM 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 shrinking. This is quite common as social networking 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 smaller and smaller place. In the end, it will become a random graph with extremely high probability. In particular, if people still keep $preferentialattachment$, the graph will exhibit a Gaussian distribution. In another word, the number of new friends are proportional to the number of friends already made, and this will keep doubling. In both situations, the network will lose the power law distribution of a scale-free network.

These cases indicate that the growth of social networking sites will ultimately lead to their loss of scale-free character. The loss of power law distribution will damage the inherent properties of a real-world social network and subsequently affect the search and spread of information and knowledge in the network, although the network may have smaller average paths and larger clustering coefficients than real-world network. Worse still, the socialising footprints and the deformation of the topology could compromise the trust and cause privacy leaking. Old members will leave the network because they feel unfamiliar and unsafe. New users will not join in because the sites lose attraction. Thus, the networks will eventually stop growing and start shrinking.

In fact, even before the number is over 150, these processes have been taking place in some social networking sites. A good example is Friendster, as we discussed above. The existence of fakesters and acquaintances greatly devalued the friendship links. People do not trust friends' contacts as in real world because there are huge amount of loose acquaintances. As the company concerned the problem of fakes, they attempted to remove them manually. This develops tension between the site and the users, even leading to some rebellions. Because of this, many members left the network, others tend not to join the network due to its bad reputation.

Most social networking sites recognise such problems and attempt to eliminate the socialising footprints. Examples include rating systems and friendship management. These measures may slow down the process of randomising the network, but far from maintaining the shape of scale-free characteristics. This is because the fundamental idea lying on these measures is to keep *preferential attachment* by removing socialising footprints. If such measures carried out in a manual fashion, people may seek to boost their "degrees", the key factor of *preferential attachment*, by manipulating their footprints. Thus, the networks will still lose the social network topology. They are still in the risk of losing trust and leaking privacy.

Chapter 4

The RealSpace Social Network

Most issues confronting social networking sites come from the fact that they are modelling people's dynamic ever-changing real-world relationships in a static model. The static model holds an implicitly stationary view of relationship formation in which connections, once formed, were permanent – thus entailing zero maintenance cost[9]. The static model damages the properties and topology of real-world social network. Thus, we propose RealSpace, a social networking platform based on dynamic social network model. Our goal is to establish an online social network by capturing the connections in the real world. This dynamic network model is also capable of eliminating relationship footprints automatically. We develop two types of search algorithms for resource locating: exhaustive search and decentralised search. Exhaustive search is fast and expensive, but may not be able to find the item of interest; on the other hand, decentralised search is relatively efficient and simple, and it can usually locate the target in a few steps.

4.1 Dynamic Social Network

In real-world social network, people and their relationships 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[6]. Common methods of Social Network Analysis (SNA) have tended toward static analysis of role relationships, based on the accumulated data over the entire time of observation. On the other hand, social network dynamics have historically been of interest, but data were limited[37][10]. Thanks to the rise of Internet and email usage, the study of dynamics and evolutions of social relationships has attracted more attention due to the availability of data. It has been pointed out that *dynamics* have two meanings that are worth distinguishing[38]. First, in the macro level, *dynamics* refer to the evolving structure of the network itself, the making and breaking of network

ties. Second, in the micro level, it refers to the dynamics of the individuals in the network. Most research demonstrates that the static topology does not capture the dynamics of social networks[8][9][18][25]. Latest study on social networking sites such as Cyworld[1], MySpace and Orkut confirms that online networks reflect real-life social networks, which is large-scaled and extremely dynamic[1]. Based on these investigations, RealSpace introduces dynamic analysis to online social networks.

Active Contact

We establish new connections whilst communicating 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 1. Then as time passes, strength S decays exponentially if they do not exchange information[19]:

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

where k is an adjustable parameter of the model and is set to be 0.001 in our case. Figure 4.1 shows the change of **connections strength** over time.

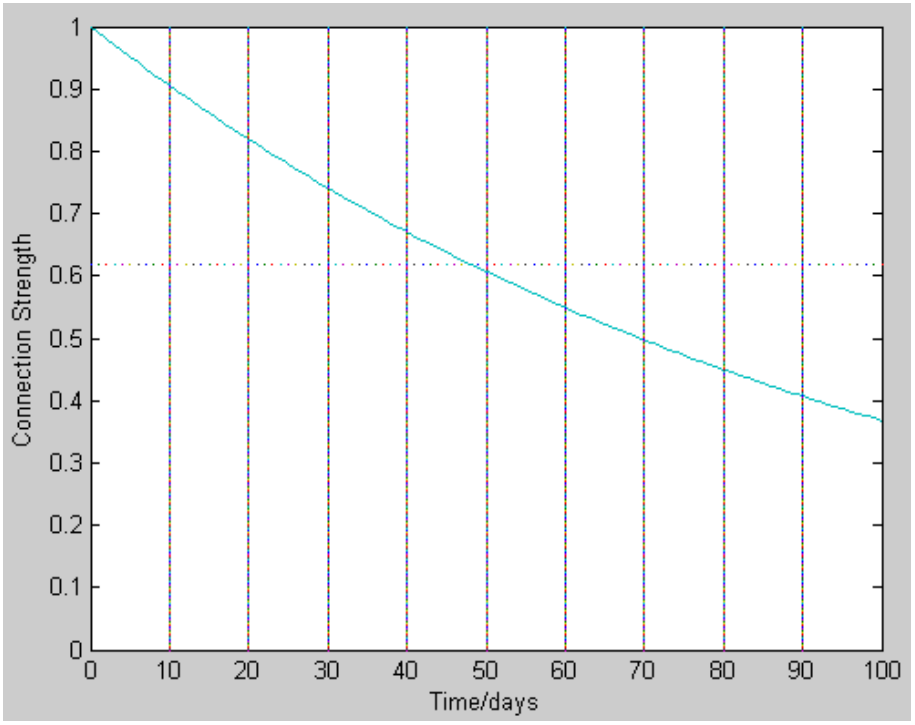

FIGURE 4.1: The Connection Strength - Time Diagram

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

[1]Cyworld: http://www.cyworld.com

Given the concept of connection strength, we introduce the definition of **active contact**. The person v is an active contact of the person u if $S(uv) > T$ while T is the active threshold which we set it to be T=0.618 for our convenience. Thus, in the diagram above, person v is not an active contact of person u until after 50 days if they do not communicate after the first contact. After 50 days, v becomes an *inactive contact* of u.

Justification of Active Contact

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 is similar to the one we employed to describe the strength of connection[12]. This 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 on maintaining the existing relationships which we do cherish. The application of active contact can effectively exclude casual acquaintances as analysed 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 list 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 connection are entirely based on the frequency of communication and interaction, which conforms to our previous analysis. Active contacts is also useful in distinguishing real users from fakesters. Real users do not normally communicate with fakesters. Thus, fakesters are often in the status of being inactive. If real users do communicate with fakesters, then fakesters turn to active. This is the case where real users use fakesters as their online personas.

4.2 Social-Connectivity Based Exhaustive Search

Most social networking 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, as discussed in the previous chapter. However, we have pointed out that the algorithm is easily subject to abuse by users as over-rating or under-rating. This is true particularly when the users see the benefit 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 target individuals' degree. Thus, better connection will attract more friends. Therefore, we develop a search algorithm based on social connectivity to improve the search quality.

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

$$C(u) = \sum_{i \in N_{\mathcal{N}}(v)} P(i)C(v_i) \tag{4.1}$$

where v_i is the i^{th} active contact of u and $P(i)$ is the weight of the connection between u and v.

Social connectivity is essentially eigenvector centrality using only active connection. In SNA, eigenvector centrality has long been used to signify the importance of a node in the network[36]. The fact 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.

4.3 Social-Distance Based Decentralised Search Algorithm

Although exhaustive centralised search is very fast and it takes only one step to get the result, it is very expensive. The algorithm needs to look up all the items in the database when answering a query. Secondly, the index table has to be updated regularly to reflect the change of the data. This introduces maintenance cost to the algorithm. All the cost will increase as the database grows. Finally, the algorithm may simply not be able to find the items of interest. Even with a sophisticated ranking mechanism, people still have to check the returned results one by one. If the item of interest is not in the top list, then one could only issue another query and keep clicking the mouse. Exhaustive search can really make one exhaustive. Yet it can not guarantee a right result.

To overcome the weakness of exhaustive search, we develop a decentralised search algorithm based on the notion of social distance. The algorithm is derived from the social-distance tree model as discussed in chapter 2 with some modifications. There are two steps in the algorithm:

(1) If one knows the answer to the query or knows a friend who understands the answer, he will reply the query or put the query to his friend. The answer will be returned directly to the original sender. The spreading of query stops once the sender confirms the answer.
(2) Otherwise, one will consider his friend whom he believes is closest to the answer. A two-dimension table is constructed to help search the relevant forwarder. In case there are more than one candidate in the group, a closeness-based ranking is employed to rank the people. In particular, if a candidate would like to share his or her social table, one

can immediately view his friend's table, possibly with some restrictions or some form of permission. The transferrable table is illustrated in Figure 4.2. $H_1, H_2, H_3, \cdots, H_n$ are social categories such as geographical locations, occupations and hobbies. Every grid will show up a group of candidates who belong to both categories.

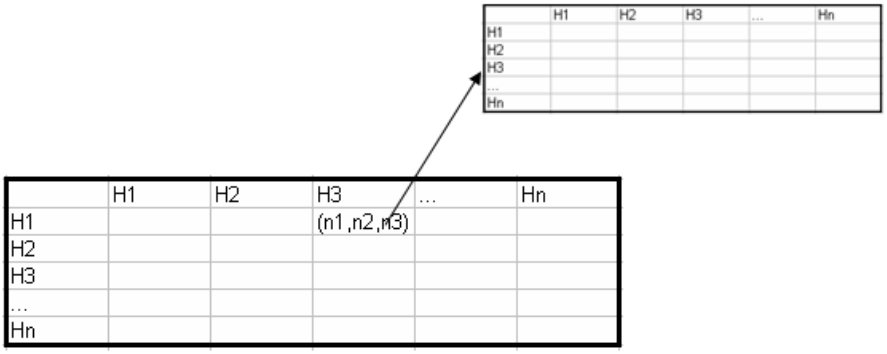

FIGURE 4.2: The Direct Query of Friend's Friend

The transferrable table is not mentioned in any previous research. It is proposed as a novel idea in our decentralised algorithm. The idea has the origin in the social networking sites where people can easily communicate and share information with multiple friends in various channels.

The closeness mentioned in the algorithm is calculated as follows:

$$C = \frac{1}{\sum_i d(u,v)} \tag{4.2}$$

$d(u,v)$ is the shortest network distance between u and v. It has been mentioned in chapter 2 that the social search coupled with node degree yields better performance. We attempt to further improve the performance by using closeness centrality. The rationale is this: the potential forwarder is expected to be most close to the final target. Thus, the closeness of the candidate is theoretically more important than his degree.

In theory, the algorithm can reach the target within $O(logn)$ steps, compared with $O(n)$ in centralised algorithm. Thus, unlike the centralised search, which tends to take more time as the network grows, the complexity of the decentralised algorithm is relatively small as the network grows. For example, a network with 1 million members takes about 6 steps and 1 billion takes 9 steps. Another concern of the algorithm is the response time, since one may not be able to reply or forward the system if he is not currently using the system. Therefore, it is also very useful to design a flag to signify the availability of the user. Sender can use the flag to find people who are currently online and are able to provide the service.

Chapter 5

System Implementation

We have developed a prototype for the RealSpace System. The prototype is a social networking site with basic functionalities and are written in PHP. We will first give a high level architecture overview of the system. Then, the structure of component modules is shown to provide some details about the architecture. More details about data schema and applications will also be discussed in the later sections.

5.1 Architecture Overview

Figure 5.1 represents the high level architecture of the RealSpace platform. In the centre of the graph are the communication and interaction between registered users that the system is going to capture. The second layer is the abstraction of the relationships using dynamic social network model. The third layer are essential utilities of the platform, such as impression management tools (profile editors), communication facilities and people search engine. 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.

5.2 System Structure

Figure 5.2 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 mainly reads 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

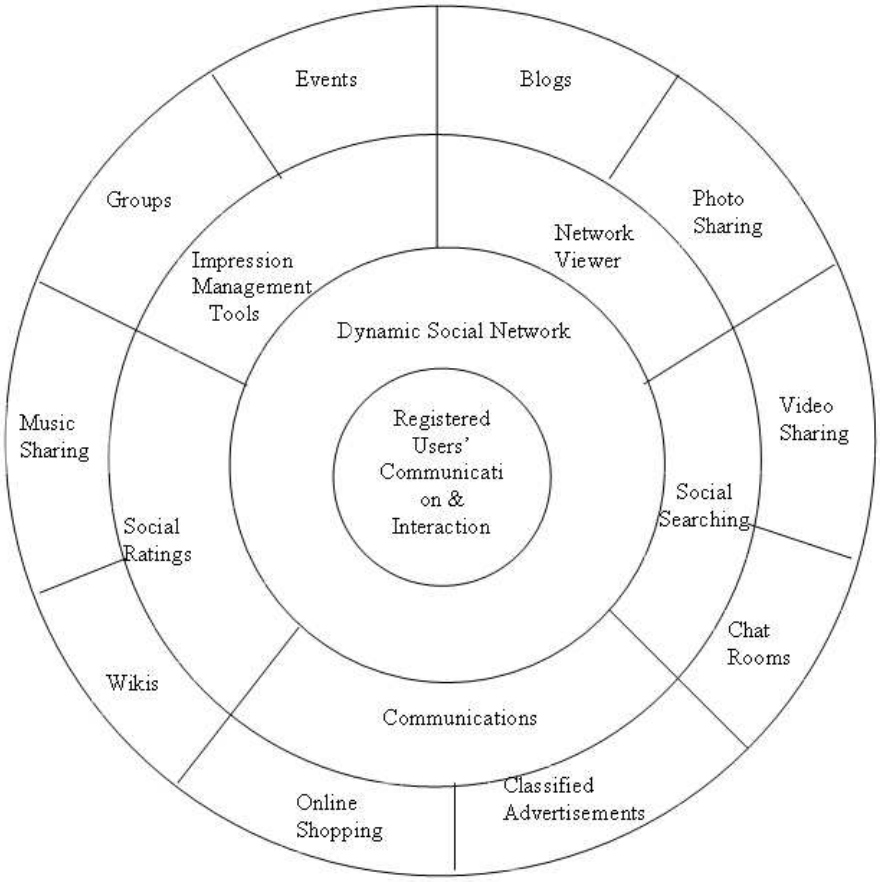

FIGURE 5.1: RealSpace Architecture

checker for updating the record. Diamond boxes are auxiliary units that aims to improve the performance of the major components. The activity checker, in particular, is responsible for refreshing the real connections between the people. 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.

5.2.1 Database Schema

The database is the soul and heart of the system. Unlike data repository of Web search engine, which generates the data by crawling the Web, a database of social networking site captures the data input by the users. An HTML document may just include creation time, headline, metadata and full text while the record of a person will have many more dimensions of information, which can be very flexibile. Thus, a detailed schema is necessary to provide rich descriptions of the data. In Figure 5.3, for example, the table *Users* include $uid, firstname, surname, email, occupation$, etc. These information are either input by new members when they register or supplemented by existing members when if they need to. The data will be used by almost all other modules and therefore

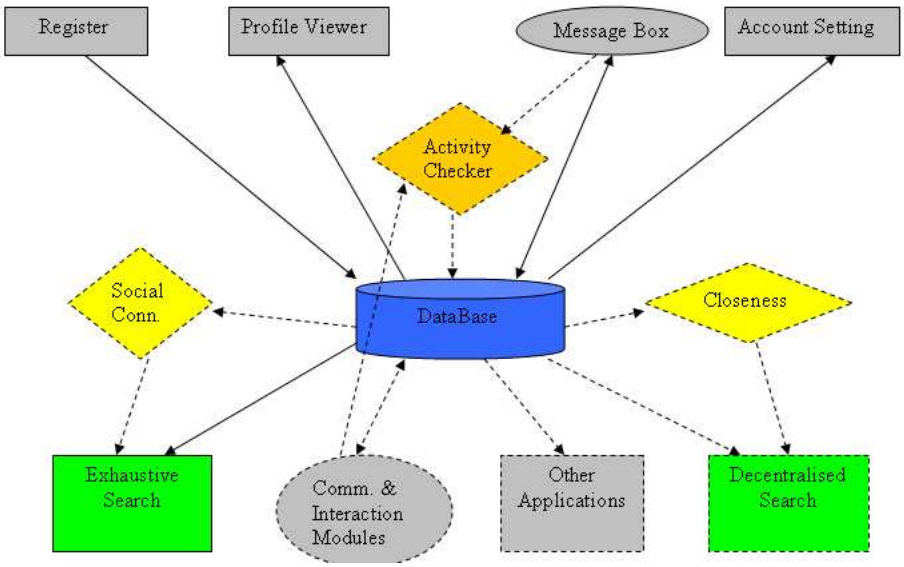

FIGURE 5.2: Major Component Modules

we write a query interface that specifically to insert and retrieve the data from the table. The format of email will be checked before entering the table. Initially, this included 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, due to the reasons discussed in the following section. 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. The field of *fuid* represents the user who initiates the friend request; *type* describes the category of friendship; *reject* indicates how many times user *tuid* has rejected the invitation. User *fuid* is banned if his request has been rejected by the same user for more than three times. The table is critical in constructing the relationships in the network. The format appears to be a directional edge in our table but it will be considered as undirectional in most applications.

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 5.3: Database Schema

5.2.2 Exhaustive Searcher

RealSpace's exhaustive searcher helps users to find users in the database. There are two interfaces with the searcher: a general search interface and an elaborate search interface.

On the general search interface, one 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. On the elaborate search interface, as shown in Figure 5.4, one can make a query by specifying the details of the person such as his occupation, hobbies and residence. The searcher will make an intersection rather than union operation on this criteria such 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.

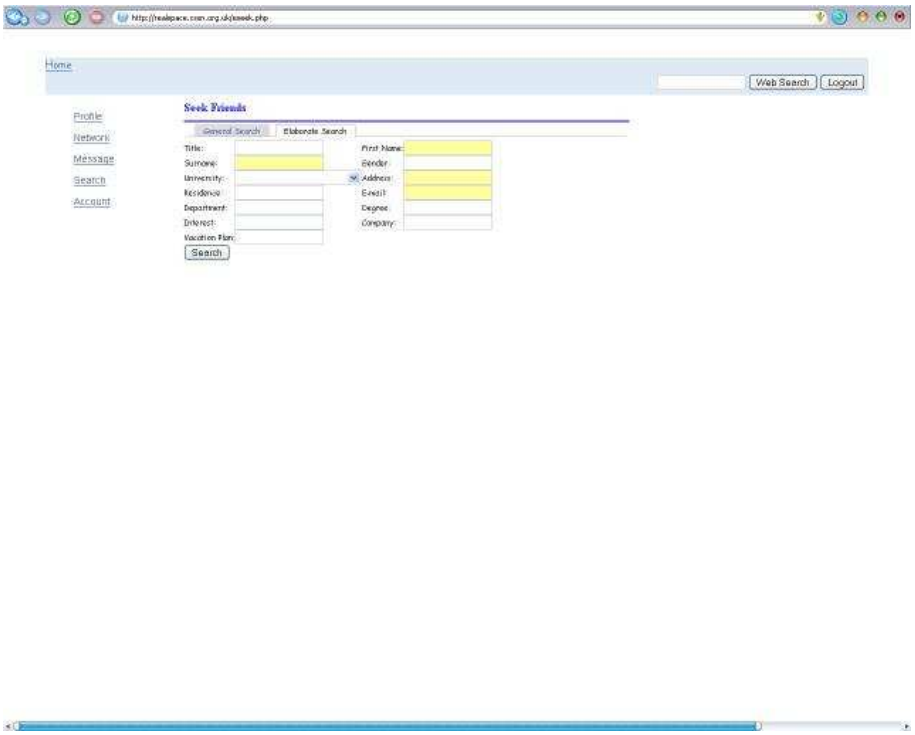

FIGURE 5.4: The Interface of Elaborate Search

5.2.3 Validating Registered Users

The majority of social networking 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. The disadvantage is that there are less coherence and integration in the network. Users may feel less committed to the relationships which are acquired through the websites. Some networking sites require a certain form of identifier or invitation. Orkut and Facebook were examples of these kinds, though that requirement

is now abolished due to the commercial interests. In 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 and 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. Validation of new registration should be simplified. We believe the issues such as loose acquaintances, fakesters and trust can be better addressed by using dynamic network. Setting up the policy for validating registration is not a long-term solution.

5.2.4 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 as 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 networking sites provide privacy settings for users to control the information flow. However, commercial networking 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 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 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.

Chapter 6

Future Work

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 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.

Integration of Long-Range Rewiring in BA Model

Our goal is to design a scale-free online social network. The idea is conceived according to our model which predicts the problem of growth constraint in many social networking sites. Our model supplements the BA model with the key element of Kleinberg's model, that is, *long-range* shortcuts in power distribution. The model is used to explain the growth of social networking 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 and power law 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 argue 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.

Managing the Range of Connection Strength

So far, our system can only determine two types of relationships: acquaintances and close friends. The system simply 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 networking sites is to retain history of all these relationships, allowing users to accumulate and utilise the contact resources without memorising them. Thus, a useful social networking 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 distribution. They hypothesis should be further scrutinised against data from social networking sites and should be consistent with existing models.

Improving Decentralised Search Algorithm

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 choice. 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.

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.

Bibliography

[1] 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.

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

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

[4] M. Andrews. Decoding myspace. *U.S. News & World Report*, 18 Sept. 2006.

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

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

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

[8] D. Braha and Y. Bar-Yam. From centrality to temporary fame: dynamic centrality in complex networks. *Comlexity*, 12:59–36, 2006.

[9] 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*, June 2006.

[10] P. Doreian and F. N. Stokman. *Evolution of social networks*. New York, Gordon and Breach, 1997.

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

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

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

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

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

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

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

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

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

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

[21] H. A. Kautz, B. Selman, and M. A. Shah. The hidden Web. *AI Magazine*, 18(2): 27–36, 1997.

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

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

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

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

[26] 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.

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

[28] 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.

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

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

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

[32] P. Resnick, E. Zeckhauser, E. Friedman, and K. Kuwabara. Reputation systems. *Communications of the ACM*, 43, 2000.

[33] O. Simsek and D. Jensen. Decentralized search in networks using homophily and degree disparity. 2005.

[34] 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.

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

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

[37] S. Wasserman and K. Faust. *Social network analysis*. New York, Cambridge University Press, 1999.

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

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

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