View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2009 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 8-3-2009

PRIVACY IN SOCIAL NETWORKS: A COMPARATIVE STUDY

Shan Chen
Innovation and Enterprise Research Laboratory,UTS Australia, shanc@it.uts.edu.au

Mary-Anne Williams
University of Technology, Sydney, mary-anne@it.uts.edu.au

Follow this and additional works at: http://aisel.aisnet.org/pacis2009

Recommended Citation

Chen, Shan and Williams, Mary-Anne, "PRIVACY IN SOCIAL NETWORKS: A COMPARATIVE STUDY" (2009). *PACIS 2009 Proceedings*. 81.
http://aisel.aisnet.org/pacis2009/81

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2009 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

PRIVACY IN SOCIAL NETWORKS: A COMPARATIVE STUDY

Shan Chen
Innovation and Enterprise Research Laboratory
University of Technology, Sydney
NSW Australia 2007
shanc@it.uts.edu.au

Mary-Anne Williams
Innovation and Enterprise Research Laboratory
University of Technology, Sydney
NSW Australia 2007
mary-anne@it.uts.edu.au

Abstract

Social networks provide unprecedented opportunity for individuals and organizations to share information. At the same time they present significant challenges to privacy that left unaddressed will stifle information sharing and innovation. In this paper we analyse four different prototypical existing social networks, and identify key problems that arise for a privacy-by-design approach to the development of a new breed of social networks.

Keywords: Privacy, Social Network, Relationship.

1 INTRODUCTION

Social networking has been an important activity within human society for millennium. Online *social network* (SN) sites provide unprecedented opportunity for individuals and organizations to share information. They provide a platform to facilitate communication, exchange and sharing of information among users and third parities. Due to the Internet, socialization on SN sites is free of geography and time restrictions. These advantages over off-line social networking have quickly attracted an enormous number of users. As a result, online SN sites have rapidly increased in number, size and scope. To attract users in a competitive market, SN sites offer innovative services to facilitate socialization. However, these services also introduce privacy infringement risks to their users as a result of the nature of information exchange and sharing on the Internet.

The underlying problem that can lead to privacy infringements is the user's lack of *information flow control* – i.e., the power to restrict information distribution. Generally, privacy infringements occur due to the user not being made aware of, or being able to monitor and control "who can see what". Specifically, this problem can be decomposed into five *operational privacy problems* (OPP): i) *what* information will be disclosed? ii) *who* can the information be shared with? iii) *where* the information would or could reach? iv) *when* the information would or could reach who and where? and v) *how* the information would or could be used? While the first two problems have been noted and are being addressed, the other three have not received sufficient attention. For example, although Facebook (www.facebook.com) has offered different levels of privacy control, the implementation of the transitive relation Friend of a Friend (FOAF) - i.e., a FOAF is a friend - is still problematic, e.g., it can easily cause information flowing to inappropriate or unintended parties. Unfortunately, FOAF has been widely accepted as a de facto standard vocabulary for representing online social networks. While it is successful in terms of *use*, from the service provider's perspective; and *networking*, from the user's perspective, the privacy issue it raises is still largely ignored – the underlying problem of the privacy issue, i.e., why users cannot control their information – i.e., the "why" problem.

In order to explore this fundamental problem we look at its philosophical roots in the physical world - i.e., *relationship-based social interactions* drive the development of human society. Clearly, the "why" problem reveals that current business model of social networks fail to meet users' needs. This failure suggests privacy-by-design needs to be a requirement for next generation of SN applications (as opposed to current models fixing the problem after applications go live). To understand the privacy problem in online social networking, this paper analyses four different prototypical existing SN sites. Using a comparative study, a number of key privacy related problems are identified, and consequently design guidelines are proposed for privacy-aware SN applications.

The rest of the paper is organized as follows: section 2 shows a motivating scenario which highlights the privacy challenges that can naturally arise, section 3 studies the concept of social network and related formal definitions that support online social networks, section 4 reviews four prototypical SN sites in relation to the challenges identified. Based on the study above, section 5 presents a comparative analysis of the prototypical SN sites from a privacy perspective. Section 6 draws a conclusion on the findings and looks at future works.

2 MOTIVATION

This section describes a scenario based on the transfer of information from a financial professional's off-line to online social network. The scenario highlights key privacy challenges that arise with respect to her requirements and expectations.

2.1 Scenario

Helen is a financial professional who aims to keep her personal and professional life separately. So, Helen has two social networks: PSN for her personal friends and WSN for all her professional contacts. The PSN has personal contacts *Matt*, *Jeff*, *Phoebe* and *Greg* – *Helen* knows *G*reg from her coursemate Matt's housemate *Jeff* who is a friend of *Greg's* wife *Phoebe*; whereas the WSN has professional contacts *Anna*, *David* and *Kathy– Helen* works with *Anna* and *David* in the same lab and she knows *Kathy* from a work related online discussion group.

Helen believes the PSN and the WSN are completely disjoint. However, a few days later, she discovers that: i) *Anna* became *Matt*'s friend (off-line) and joined the PSN by *Matt*'s invitation, ii) *David* invites their boss (of the lab) *Bill* to join the WSN for work purposes, iii) *Bill* establishes both a personal and professional relationship (off-line) with *Greg* and invites him to join the WSN, and iv) *Greg* found his lost-contact ex-colleague *Kathy* on the network. When *Helen* is aware of all these changes, lots of her personal information has already been made known to her contacts in the WSN, and work-related information has also flowed into the PSN. Figure 1. shows the evolution of the PSN and the WSN.

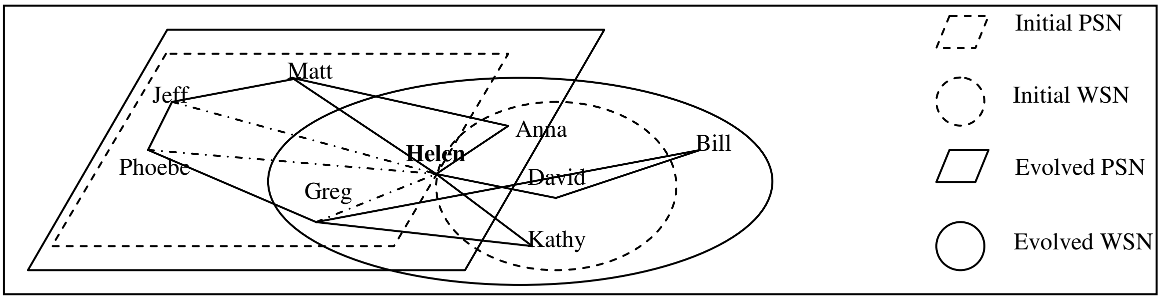


Figure 1. Evolution of Helen's PSN and WSN

2.2 Requirements

Helen wants to have control over her personal information in the PSN. She also wants to keep her WSN strictly professional, i.e., no personal information should enter the network.

On the PSN, *Helen* sees: "all coursemates are friends", "all housemates are friends", and "all friend's friends, and friend's friend's friends are friends". She only wants to share course-related information with her coursemates and household related information with housemates. She shares different personal information with different friends, e.g., her coursemates know her school email address and skype id, housemates know her hotmail address and msn id, etc. She requires anyone who she shares her information with not to distribute the information to others without her consent. This implies that if anyone on her PSN wants to invite others to join the network then they need to get her consent.

On her WSN, she wants to keep the network for professional related socialization. Anyone joining the network must agree not to share any information from the network with any non-professional-related persons. Similar to the PSN, this also implies that if anyone on her WSN wants to invite others to join the network they will need to get her consent.

In the off-line world, *Helen* and *Jeff*'s friendship tends to be developed into a boyfriend and girlfriend relationship. As a result, the online information they are sharing is getting more and more personal. *Helen* does not want others to know about this relationship for the time being and she also asks *Jeff* to keep the information between them.

Helen's requirements and expectations can be summarised as having abilities to: i) deal with different kinds of friends like coursemate, housemate, etc.; ii) recognise friends within 3 degrees away (i.e., restrict FOAF within 3 links); iii) control who can join the network; iv) control who can know which of her relationship to others in the network; and v) be aware of any changes or intentions of changes.

2.3 Privacy Challenges

Privacy challenges arise from *Helen*'s requirements and expectations (listed in section 2.2) are as follows:

- With regard to (i), the problem concerns the user's ability to control *access level*, i.e., the range of privacy settings for users to control their information by granting different levels of access for their social contacts. This problem of access level reflects the problem of *relationship granularity* - if we look at the problem from a different angle, the ability of dealing with different kinds of friends implies the need of various relationships at different level of granularity.

- With regard to (ii), the problem concerns the user's ability to control *information flow*. There are two fundamental sub-problems:
  - *Relationship granularity* - This problem is closely related to the problem of access level in (i), since the lack in recognizing the need of multiple types of social interactions can easily lead to information flows to inappropriate parties.
  - *Connection degree* - This problem concerns how tolerant an individual's network can be in terms of information distribution. When multi-granularity is developed, the tolerance involves granularity, e.g., different tolerance for different types of relationships.
- With regard to (iii), the problem concerns the user's ability to control his *connection space*, which has a large dependency on the relationship types supported by the SN sites. On the other hand, the user needs to have the central power on his own social network such that an accepted joining invitation will not be approved without an approval from the central user.
- With regard to (iv), the problem concerns any of or any combination of the problems (i)–(iii). With regards to (v), the problem concerns the capacity of the system's notification feature. Email notification is the primary feature in these four representative social networks. Thus, measuring the system capacity is mainly on what and who can be notified.

From a privacy perspective, all these problems are about "who can know what". The 5 OPP questions provide a good grounding for this problem. The "why" problem raises a *relationship privacy* issue whose fundamental problem is *relationship granularity*. Sufficient granularity of the supported relationships allows users to better utilize the capacity of their (and others) connection spaces to develop them to best meet their own needs. Such development includes modelling access levels and connection degrees. Effective development of connection space can facilitate learning necessary notification elements and improve designation of notification system, which in turn will drive and advance the development of connection space. In summary, relationship space provides a platform for developing access level and connection degree, operating connection space and designing notification elements. In this light, we regard the relationship privacy the primary privacy problem, and the relationship space and connection space the primary problem domain. In the next section we study how social networks are modelled and the potential of the existing models to support addressing the problems we have identified.

3 SOCIAL NETWORKS AND FORMAL DEFINITIONS

The term "social network" was first coined by sociologist Barnes (1954) who defines a social network as a group of around 100 to 150 people. Nadel (1957) developed the term to the concept that underlines the notion of "role" as the foundation of social lives. Since then, the concept of *social network* has been developed as a structure (in contrast to "content"), with a emphasis of connections between social entities. As Wasserman and Faust (1994) have pointed out, the concept of *network* implies the essentials of *ties* between or among its members. Thus, fundamental to a social network is ties that connect social entities and the social implication of these ties.

Mathematical models have been developed for social network theory. Various notation schemas have been studied, mainly graph theoretic, sociometric and algebraic that can be adapted to represent a wide range of social networks. *Graph notion* is straightforward, it provides an elementary way to represent social entities and relations among them. *Sociometric notation* is the most common studied in social network literature. It uses a sociomatrix to represent related pairs of entities. *Algebraic notation* is used to study role structures, i.e., multiple relations between entities. Although these three schemes overlap to some extent, Freeman (1989) views a social network as: $<S, G_d, X>$, where i) $S=<N,L>$, N is a set of nodes and L is a set of directed ties connecting nodes in N, ii) G_d is a directed graph (sociogram) generated from S, and iii) X is a sociomatrix $N \times N$. Wasserman and Faust (1994) have noted that social entities attributes are not easily captured by just using these concepts. As a solution, they introduced a new matrix, A, defined by $|N| \times$ (number of attributes). In this light, Freeman's definition is extended to $<S,G_d,X,A>$ (*Def1*).

Recent development of social networks has been extended from off-line to on-line, for building recommenders, socialization, information sharing, collaboration, etc. For example, Liu *et al.* (2008) have studied the problem of privacy-preserving data analysis over graphs and networks. By nature of the problem, they have modeled a social network as a graph $G=(V_G,E_G)$ where vertices $V_G=\{v_1,...,v_n\}$ denote individuals and edges $E_G=\{(v_i,v_j)|v_i,v_j \in V_G, i \neq j, 1 \leq i,j \leq n\}$ denote social relationships among individuals.

In an attempt to discover social networks for personalized mobile services, Jung *et al.* (2008) have defined a social network as $S=<N,A>$ where N is a set of participants $\{n_1,...,n_{|S|}\}$ and A denotes a set of relations between the participants, represented by an adjacency matrix consists of binary values where 1 denote the existence of a relation and 0 otherwise. However, there are multiplex social networks beyond simple binary relations can capture. The authors use an additional component, C, to specify multiplex social relations. By attaching C to A, a multiplex social network S^+ is then defined as $S^+=<N,A,C>$ (*Def2*).

Zhou and Pei (2008) have modelled a social network as a graph $G=(V,E,L,\mathcal{L})$ (*Def3*), where V denote a set of vertices, E is a set of edges such that $E \subseteq V \times V$, L a set of labels organized in a hierarchy, and $\mathcal{L} : V \rightarrow L$ assigns each vertex a label. The novel idea of this model is the L can be utilized to generalize the labels to anonymize the neighbourhoods of vertices - a way to prevent neighborhood attacks to achieve the goal of preserving individuals' privacy.

Compared to approaches that use adjacency matrixes, Carminati *et al.* (2007) have modelled social networks in a higher abstraction level using relationship types. By introducing a trust level, they define a social network in a 5-tuple such that, $SN=(V_{SN},E_{SN},RT_{SN},T_{SN},\Phi_{ESN})$ (*Def4*), where V_{SN} and $E_{SN} \subseteq [V_{SN}]^2$ are nodes and edges of a digraph (V_{SN},E_{SN}), RT_{SN} is a set of relationship types, T_{SN} is a set of trust levels, and Φ_{SN}: $E_{SN} \rightarrow RT_{SN} \times T_{SN}$ assigns each node in E_{SN} a relationship type in RT_{SN} and a trust level in T_{SN}.

It can be seen that, different purpose social networks have different problems, requiring different definitions of social networks. Common to these definitions is a graph that consists of nodes connected by edges. However, a graph is only a structure that represents social connections syntactically. The semantic aspect of the social connections needs to be captured in a social network model. For example, if the social network allows multiple relationships co-exist between two parties, can the definition capture such semantics? In case of relationship privacy, how can the definition facilitate "hiding" a relationship or part of a relationship's information? In section 5 we examine where these formal definitions meet the four representative social network models that presented in the next section.

4 PROTOTYPICAL SOCIAL NETWORKS

There are over 250 websites featuring social networking that provide services for users to build and browse lists of contacts. These websites allow users to communicate with each other for some purpose (e.g., blogging, business, dating, pets, photos, religious, social/entertainment, etc.) In this paper we do not aim to provide comprehensive categories for existing SN sites. Instead, we are interested in platforms that support socialization for the general public. For analysis purposes we look at the most popular ones that i) provide a *general-purpose* platform for socialization, ii) have higher *popularity* or provide *innovative* platforms, and iii) provide *free-access* to the public. Then, based on their *service goals/missions*, we have selected four key representatives:

- *Facebook* (www.facebook.com) - the most popular general-purpose social network that has the fastest growth profile in number and size.
- *LinkedIn* (www.linkedin.com) - the most popular business/professional-focused platform.
- *Pulse* (www.plaxo.com) - the first to aim at real world relationships.
- *Chi.mp* (http://chi.mp) - the first to provide individual-centric platforms, a new member of social network family that claims to be the next generation social networking site.

In the following we describe these four representative examples and their features in relation to the capability to address the problems identified in section 2.

Facebook provides a platform for users to get in touch with friends and contacts. It also offers opportunities for users to meet and interact with an extraordinarily expansive universe of new people to create new relationships and communities. Its free access and open environment advance individuals and organizations a wide range of socialization. Features developed to support socialization are mainly: i) *Wall*, which offers a place for friends to post message to the user; ii) *Pokes*, which allows users to "poke" to each other to gain their attention; iii) *Status*, which enables users to inform their contacts of their whereabouts and actions; and iv) *News Feed*, which automatically share users activities (e.g., profile changes, Wall posts and newly added friends) with their friends by the user's consent (via setting). On the other hand, Facebook provides a platform for software developers to create applications that can benefit users. Applications launched include i) *Photos* to upload unlimited photos, and allow putting comments and tags to share with family and friends; ii) *Videos* to share homemade video; iii) *Groups* to create sub-networks for better socialization; iv) *Events* to inform friends about upcoming events; v) *Notes* to share stuff with friends; vi) *Links* to provide links to friends post; and ix) *Gifts* to send virtual gifts to each other.

In terms of relationships, Facebook provides limited capability for users to develop their connection space. Relationships recognized in Facebook are Friends, which can be "Friends of Friends"(FOAF), "Only Friends" or "Only Me". Users can also maintain relationships with others through "Networks" or defined Groups at an abstraction level, where within the group/network relationship between individuals (member of network/group) can only be "Friends".

LinkedIn was designed for professional networking with a goal to facilitate connection between people for work-related purpose. With an emphasis on building a reputation and connecting to employment and business, LinkedIn networks consist of business *connections* only. Users build networks through *introductions* or *referrals*, within 3 degrees away. Users can build groups, define who has access to their material and view their profile. They can search for companies in their network, as well as allowing others to search them.

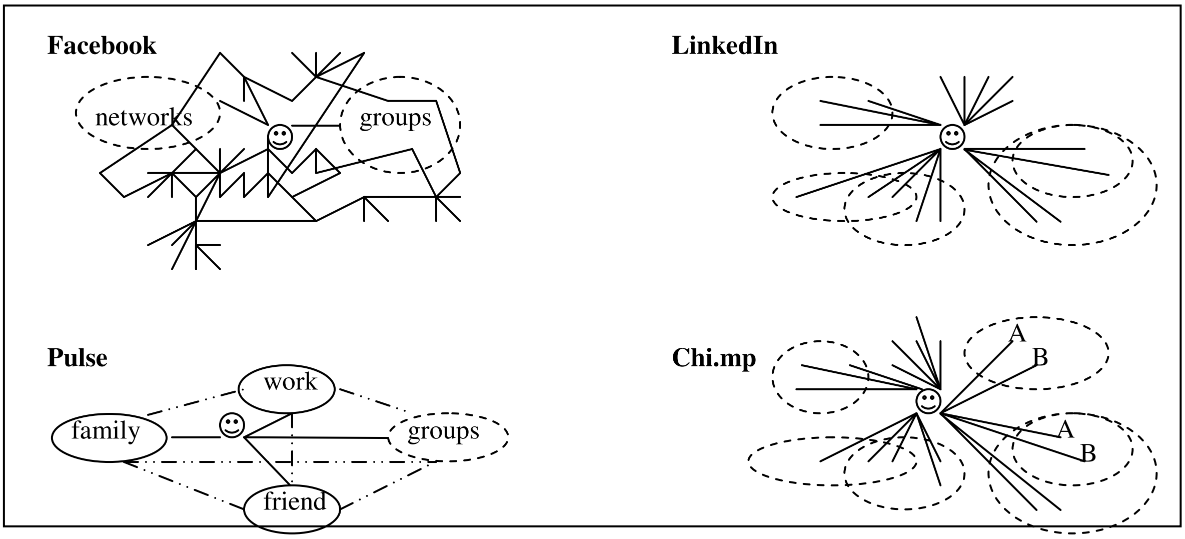
Pulse is a social networking service of Plaxo (www.plaxo.com). Unlike Facebook that measures users' popularity on the number of *friends* they are connecting, Pulse aims at real world relationships and encourages users to stay in touch with people they actually know. It commits to real relationships by endorsements – connections are built via email consents at both ends. By this approach, users can only establish connections with those they have email contact.

Pulse creates three social spaces for users: Business Network, Friends and Family Member. In addition to these social groups, users can create customized groups in which contacts are connected on the relation of "Business Network", "Friends" and/or "Family Member". Any changes on the relation, i.e., connect/disconnect and connection types, require endorsements at both ends of the relation. As a result, connection degree is always 1.

Chi.mp stands for "Content Hub & Identity Management Platform". Different from other SN sites, Chi.mp provides each user a free .mp domain which serves as an OpenID to house and share online content. It allows users to i) import their online content (e.g., blog entries, photos, video, tweet, etc.) scattered across number of sites, and ii) import and merge contacts scattered across platforms (e.g., Outlook, Address Book) and Web services and social networks into their .mp domain (OpenID) to centralize their identities in one place. It features *personas* for identity management. By creating different personas for different content and assigning them to different contacts, the central user obtains control over "who sees what" on their own domain. Chi.mp provides three default personas: Public, Work and Friend. Users can create own personas as many as they want.

In summary, these four SN sites provide users a platform to build social networks online. Common to these social networks is their profile-based approach. While in the first three SN sites each user is a node of the central network and networks of a user are sub-networks of the central network, Chi.mp allocates each user a domain to build a network with the user as the central node. Connection types (relationships) between contacts are supported in different degree of granularity on different SN sites.

LinkedIn, Pulse and Facebook provide fixed options with customized groups, whereas Chi.mp offers flexibility to users to define relationships for their own needs. As such, connection degree are various on these four SN sites. Figure 2. shows topologies of a user's network in the four SN sites.



*in Chi.mp circles indicate resource, whereas they indicate groups/networks in the other three SN sites

Figure 2. Topology of a user's network in four representative SN sites

5 COMPARATIVE ANALYSIS

This section compares the four SN sites in relation to privacy issues including i) relationship space that provides fundamental support to carry out socialization, operate information and control privacy; ii) user privacy control that are features empower users to control their information; and iii) formal definitions that are used to support system implementation. In each sub-section, an introduction of concepts used for the analysis is presented, followed by a comparative study on the four SN sites. Based on the comparison results, several key findings are presented to support future work for a privacy-by-design approach in the development of a new breed of social networks.

5.1 Relationship Space

Relationship space is a conceptual space that holds relationship types supported by the social networks. *Connection space* accommodates all the relationships of a social network. In other words, relationships in connection space are instances of related relationship types defined in the relationship space.

The concept of *capacity* is used to measure a relationship space. The *capacity* of a relationship space has two dimensions: i) *abstraction* reflects the *level of detail* on social entities, e.g., the abstraction levels of individuals, groups, communities, organizations and networks are from the lowest to the highest; and ii) *granularity* refers to the fineness with which relationship types are categorized on a certain abstraction level. For example, in Pulse, relationship types default to Business, Friend and Family. These three types represent a higher granularity of relationships compared to the granularity in Facebook that supports only Friends relationship. If the user is a member of a group, the user's relationship to those who connect to the group can be referred to the group level. For example, Mary is a member of group Subject_IT312. Her relationship with Phoebe who is a member of group Subject_IT324 can be described as the relationship between Subject_IT312 and Subject_IT324 when talking about subjects they are studying, if the social network supports relationship between groups (e.g., ITcoursemates). It can be seen that, the lower abstraction level the higher granularity support required.

There are three dimensions that can be used to describe a connection space: direction, multiplex relationship and connection degree. *Direction* indicates symmetric/asymmetric property of a relationship. For example, the relationship between *A* and *B* is symmetric if both set the relationship of the same type under the same conditions. The relationship is asymmetric if *A* sees *B* as a friend but *B* sees *A* as a colleague. A connection is *multiplex* if more than one relationship exists between two entities. *Connection degree* is used to indicate the distance between two entities. If *A* connects to *B* who connects to *C*, then *A* is said to be 2 degrees away from *C*, i.e., the connection degree between *A* and *C* is 2.

SN Site	Abstraction: Granularity	Direction	Multiplex Relationship	Connection Degree	Privacy Control
Facebook	individual: Friends group: Network, User-defined	asymmetric (uncontrolled)	At group level	uncontrolled	Very poor
LinkedIn	individual: Connections group: User-defined	symmetric	i) indivudual level: ≤3 ii) group level	≤3	Poor
Pulse	individual: Business Network, Friend, Family Member group: Business Network, Friend, Family Member, User-defined	symmetric	At group level	1	Poor
Chi.mp	individual: Public, Work, Friend, Persona (User-defined) group: as at the individual level	asymmetric (controlled in one direction)	No	1	Good *

*can introduce great redundancy and a management overhead

Table 1. A comparison of relationship space and connection space in the four SN sites.

As shown in Table 1., the capacity of relationship space differs from social network to social network. All four SN sites provide two abstraction levels for individuals and groups. At the individual level, Facebook and LinkedIn offer the lowest granularity which allows only one type of relationship, whereas Pulse granulates relationships into three types, and Chi.mp offers the highest granularity by allowing users to define arbitrary types. At the group level, Facebook, LinkedIn and Pulse allow users to define groups. In Chi.mp, groups are not explicitly supported. The users can manually create "virtual groups" with some effort by using *personas* device as the similarity criteria.

In Facebook, relationship is limited to Friends. Consider an example: user *A* accepts user *C* as a FOAF friend. However, *C* does not accept any FOAF thus *A* is not a friend of *C*. As a result the relationship between *A* and *C* is *asymmetric*. The relationship is *uncontrolled* because *C* does not accept anyone who is more than one degree away. However, *A* and *C* share the same contact *B* who does not block *A* from knowing *C*. As a result *A* gets to see *C*. Hence, relationships in Facebook are uncontrolled and asymmetric. In Chi.mp, the user who creates the network is a central node with full control for personas that play a role as relationship connecting the central node to other nodes. Thus, relationships in Chi.mp are asymmetric and controlled in one direction. LinkedIn only acknowledges a relationship if two nodes are connected whereas Pulse permits three types of connections. A successful connection in either of these two SN sites requires approvals by both ends. Relationships in both SN sites are thus symmetric and controlled.

The sole relationship type at the individual level in Facebook and LinkedIn says only simplex relationships are allowed. On the other hand, the ability to create or join groups facilitates members of groups to refer to each other using group attributes. By this method implicit multiplex relationships at the group level can exist between group members. Pulse accommodates multiplex relationships using three fixed options. Similar to Facebook and LinkedIn, implicit multiplex relationships at group level can be customized in addition to the pre-fixed groups. While assigning one persona for each contact restricts relationships to be simplex in Chi.mp, it is possible to utilize personas to create groups manually and therefore, with some efforts, implicit multiplex relationships can be referred under a great management overhead.

FOAF in Facebook is optional, however, users do not have control over their friends' friends. Even if they do not have FOAF friends, they cannot inspect and see if their friends have FOAF friends and who are on the FOAF list. Connection degree in Facebook is uncontrollable due to the tolerance of

FOAF. LinkedIn implements Introduction within three degrees. Users can connect to FOAF within three degrees away by *introduction* and consent. In Pulse, each connection requires consent and confirmation from both ends. By this approach users in Pulse connect to each other directly, i.e., connection degree is always 1, invalidating FOAF. Chi.mp's approach of single network with a central node naturally avoids FOAF implementation, setting connection degree to 1.

In summary, the capacity of relationship space varies for different SN sites. This difference leads to the difference of connection space in the related SN sites. As a consequence users' ability to manage their personal information is supported at different and important levels. In the next sub-section a detailed comparison will be presented.

5.2 User Privacy Control

Privacy is multifaceted. Users' ability to control their privacy on a SN site can involve many dimensions, depending on the architecture of the SN site. We focus on relationship privacy with common dimensions of the four representative SN sites. These dimensions are *access level*, *consent* and *information flow*.

Privacy at *access level* concerns the user's ability to gain different levels of permissions for others to access their information. The more abstraction levels, the higher the granularity, the more access control the user will exercise, which in turn the more privacy control the user will possess. *Consent* concerns actions to establish a connection or be referred to a potential connection. *Information flow* is a crucial dimension in terms of privacy, as it concerns how far the information will travel away from its owner. The value of connection degree is used to indicate the distance. Since third parties are usually involved in the social network provider's business, they often introduce an extra layer of privacy protection and arise many issues. Once the information is published, or transferred, the user's ability to remove sensitive information needs to be considered, i.e., the power to delete their own information.

SN Site	Access Level	Consent		Information Flow Control	
		connection	referral	connection degree	3rd party
Facebook	1 (default) + number of user-defined groups	Yes	No	No	No
LinkedIn	1 (default) + number of user-defined groups	Yes	No	≤3	No
Pulse	3 (default) + number of user-defined groups	Yes	No	1	No
Chi.mp	3 (default) + number of user-defined personas	Yes	N/A	1	No

Table 2. User Privacy Control

Comparing Table 2. to Table 1., it can be seen that *access level* is supported by the abstraction and the granularity. Privacy control in this dimension can be implemented by mapping functions from abstraction and/or granularity to access level. Where access level can be set at individual and group abstraction levels, granularity is the main concern for access control. In Facebook, granularity is minimal because there are only Friends in the network. A more problematical situation is the implementation of FOAF, creating unforeseen social environments for users. Such environment can easily lead to a flood of information in the network, nourishing privacy infringement. Similar to Facebook, LinkedIn offers the lowest granularity, which simply indicates a relationship as *connection*. Such simplicity prevents users from setting up flexible access levels of control. Pulse offers more flexibility in this dimension. It fixes three connection types and allows multiple connection types between two entities, providing seven possibilities for users to establish connections. Among the four SN sites, Chi.mp provides the highest granularity since it does not put a limit on the number of personas a user can create. Users can specify as many access levels as need. The drawback with this approach is that it can introduce great redundancy and subsequently a management overhead.

These four SN sites require the user's consent to complete a connection. However, consent to being referred to other people, in the aspect of privacy, has not received sufficient attention. Facebook, LinkedIn and Pulse offer "People you may know" function to all users without any options for the user to decide if they wish to be referred to others and on what criteria.

Information flow is an extensive problem. First, it depends on the network architecture in the aspect of tolerance of connection degree. The value or the range of connection degree (Table 1.) reflects the degree of influence nodes involved inside the "wall" of the network. Second, it involves a third party issue; an issue outside the "wall". On the other hand, it has a closer relationship to the problems of access level and consent, i.e., the implementation of access control and consent functions greatly influence the control of information flow. If access level and/or consent are not comprehensive implemented, the problem of information flow can be maximally extended to include all dimensions of privacy control. It is apparent that information flow on Facebook is uncontrollable due to the lower granularity and the use of FOAF. LinkedIn empowers users to control the flow within three degrees. Pulse and Chi.mp allow only their direct connections to access information. None of the four SN sites offers full control to users in relation to third parties.

In summary, privacy control is still under-addressed in social networks. One SN site's strength might become the weakness of another one. Existing SN architectures do not provide sufficient supports for a wide range of access flexibility, tolerance for consents, responsibility for third parties involved, and power to fully control information on the network.

5.3 Formal Analysis

Modelling social networks in mathematical models is one way to support system implementation for automatic search and query-answering. This is particularly important for reasoning about privacy in online social networks. As studied in section 3, a number of formal definitions have been reported in the literature. We have selected four formal definitions that were created for different purposes, which are social network theory (Def1), personalization (Def2), neighbourhood attack resistance (Def3) and private relationship preservation (Def4).

As mentioned above, a social network consists of social entities and ties between them. In mathematical models, these elements are often represented as nodes and edges of a graph to underline the structure of networks. Examples can be found from definitions Def1, Def3 and Def4. When network content is emphasized, relationship instances are captured in an adjacency matrix like definitions Def1 and Def2.

In Def2**,** the component C captures multiplex relationship to attach to A. While it is not applicable to Chi.mp, LinkedIn and Facebook that do not support multiplex relationships, it is not clear how it can support multiplex relationships (e.g., in Pulse) mathematically. Def3 captures relationship instances in the form of labels. It organizes labels in a hierarchy to support generalization to preserve some level of relationship detail. This novel method enables generalizing relationship details at some desired levels. Moreover, a hierarchy of relationships can serve as an engine to facilitate search and query-answering. Integrating hierarchical relationships into the architecture can improve/facilitate privacy reasoning and preservation. Currently none of the four representative SN sites supports relationships in hierarchical structures.

Def4 offers a higher level abstraction model by explicitly incorporating relationship types into the model. It also dedicates to social implications in terms of *trust*. Since none of the four SN sites supports a trust function, the mapping function Φ_{ESN} maps relationship types to edges on a one-to-one basis. Multiplex relationships (e.g., in Pulse) are not supported on such one-to-one mappings. When there is only one relationship type is supported, e.g., in Facebook and LinkedIn, the mapping function makes no effort.

Def1 provides a comprehensive solution by including both structure and content in the model. In addition, it accommodates a graph, a socialmatrix and an attribute matrix to capture social entities' attributes. To a certain extent, some of the components are redundant. For example, the G_d is generated from S. However, a straightforward graph showing users their network information can assist them to understand and better organize their networks, encouraging them to be more active engaging in positive networking. Of the four SN sites, LinkedIn is the only website that provides graph visualization to the user when making connections.

It can be seen that, four definitions meet the four SN sites requirements to different degrees and ranges. Def2 has the potential to support multiplex relationships while the other three support simplex relationships only. The use of an adjacency matrix enables Def1 and Def2 to indicate directions of connections, however, they lack of ability to capture the symmetric/asymmetric property. Def2, Def3 and Def4 stress either structure or content, whereas Def1 includes both structure and content in the model. Hierarchical relationships (Def3) enable generalization towards a novel approach for preserving privacy at level of abstraction. Supplements such as graph visualisation and social entity attributes enrich semantics of network models (Def1), potentially enhancing search and query-answering. Social implication is an important component of social networks, serves as one aspect of the network semantics. Def4 has made an initiate to feature *trust* in the network model. However, social implication is far more complicated that just trust. More dimensions to support *privacy* need to be discovered and implemented.

5.4 Lessons learned

The study above highlights the problem of privacy in online social network mainly falls in the relationship space and the connection space. In other words, fundamental to the privacy issue is the problems of: i) relationship space, namely capacity which is two-fold: abstraction and granularity; and ii) connection space, namely direction, multiplex relationship and connection degree. These problems form the basis to support implementation of user privacy control on access level and information flow. On the other hand, consent and the power to delete information are important issues rely on the service provider's intentions. Mathematical models to support automatic search and query-answering implementation need to accommodate all the identified *factors* as follows:

Simplex relationship vs. multiplex relationship In the off-line world, human relationships are far more complicated than just *friends*. Multiple relationship types exist in the human society. The online social environment should simulate the off-line world to support multiplex relationships. Providing such a relationship space for social interaction can facilitate the control of access level and information flow, leading to better preservation of information privacy.

Individual level vs. network/group level Interaction on some abstraction levels, e.g., individual-to-group, group-to-group, group-to-network, etc., can avoid certain detail of information being disclosed. Subsequently, accommodating relationships at various abstraction level for interaction is required. This problem can potentially lead to the problem of relationship structure, i.e., the ability to manage hierarchical relationships and utilize them for privacy preservation.

Symmetric vs. asymmetric Relationships are often asymmetric. Better decision-making on privacy preservation requires an understanding of properties of the relationships involved.

Connection degree vs. information flow A balance between network size and information flow has a significant impact on the privacy issue. Finding out about the right balance should be considered in the model.

Formalization Mathematical models to support system development for automatic reasoning about privacy require more than just a graph structure and/or relationships at syntactical level. Semantics of networks like relationship properties and their social implications need to be accommodated in the model. Existing models do not support all these properties in a single framework. A new breed of model to address all these problems for better privacy protection is needed.

One might argue that these principles for designing social network sites should not be applied to all sites, but be contingent on the market segment. While we agree with this argument, note that our aim is at general-purpose socialization – i.e., to provide users the ability to simulate real world relationships as well as utilize facilitation in online environment that is deficient in the off-line world.

Subsequently the principles presented above are proposed as guidelines for designation of general-purpose social network sites.

6 CONCLUSION AND FUTURE WORK

The need for a privacy-by-design approach to the development of a new breed of social networks has been stimulated by the increasing privacy infringements in current online social networks and the failures to address such problems. From the four representative SN sites in our analysis, this paper has identified the conceptual problems that are relationship abstraction and granularity, direction from entity to entity with the property of symmetry-asymmetry, multiplex property and network size. These conceptual problems naturally lead to practical issues on access control and information flow control. User consent and the power to delete own information are also crucial issues. Existing models are deficient in providing an adaptive architecture from a privacy perspective.

On the other hand, addressing privacy challenges online requires software systems to offer users the ability to search and query information, reasoning about intelligent actions with respect to privacy preservation. Mathematical models are fundamental to the implementation of such systems as they help to expose the problem clearly and to find solutions to address them. Existing models are inadequate to meet the requirements identified from the problem domain described in section 5.4. Future work will take steps towards a comprehensive privacy-aware social network architecture to promote the next generation SN applications.

References

Barnes, J. A. (1954). Class and committee in a Norwegian island parish. *Human Relations*, 7, 39-58.

Carminati, B., Ferrari, E. and Perego, A., Private relationships in social networks. *ICDE Workshops 2007*: 163-171.

Freemn, L.C. (1989). Social network and the structure experiment. In Freeman, L.C., White, D.R., and Romney, A.K. (eds.), *Research Methods in Social Network Analysis*, pages 11-40. Fairfax, VA: George Mason University Press.

Jung, J. J.; Kim, K.; Lee, H. and Park, S. (2008). Are You Satisfied with Your Recommendation Service?: Discovering Social Networks for Personalized Mobile Services. *KES-AMSTA 2008*: 567-573.

Liu, K.; Das, K.; Grandison, T. and Kargupta, H. (2008). Privacy-preserving data analysis on graphs and social networks. In H. Kargupta, J. Han, P. Yu, R. Motwani, and V. Kumar, editors, *Next Generation Data Mining*. CRC Press, 2008.

Nadel, F.F. (1957). *The Theory of Social Structure*. London: Cohen ad West.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Zhou, B. and Pei, J. (2008). Preserving privacy in social networks against neighborhood attacks. *Data Engineering*, 2008.