Rhodes University

Computer Science Department

Honours Project

Identifying direct relatives of a social network site user and obtaining their personal contact details

Author:
Darryn Cull

Supervisor:
Mr. Yusuf Motara

May 29, 2015

Chapter 2

Literature Review

2.1 Introduction

Social networks have been analysed in literature for many years from the perspective of mathematicians and social scientists. However, literature about social network sites is few and far between, especially from the view of a computer scientist. In this chapter, literature about social networks will be reviewed, although more emphasis will be on literature surrounding social network sites. The four overarching topics of this literature review include social network sites, privacy, social network graphs and work relating to the design and implementation of third party applications for social network sites. It is important to establish the background knowledge needed to develop an application which uses these topics to achieve a goal. In order to identify particular members of a social network site, one must first have a good understanding of what a social network site is and how they function. It is also important to be aware of the ethics involved for example whether privacy is being broken or preserved. Additionally, one must have a working knowledge of how to use social networks: how to traverse, analyse and visualise social network graphs.

2.2 Social Network Sites

Boyd and Ellison (2007) define a social network site as "a web-based service that allows individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system." The nature of the connections made can vary from site to site.

Early work on computer-mediated communication suggested that moving from face-to-face interactions rich with context to text-based media would create an "impoverished communication environment - fraught with misunderstandings, flaming and antisocial behaviour" (Haythornthwaite, 2005). In spite of this, as new social network sites emerged and became more familiar, the use of "common and group conventions" (Haythornthwaite, 2005) have made social network sites an integral method for maintaining interpersonal connections. What makes social network sites unique, enabling users to express their social networks and make them visible to those they choose. However, this does not imply that social network sites are perfect, misunderstandings still happen and anti-social behaviour has become a regular occurrence online, from "flaming" to "trolling".

In most cases, generating a profile consists of answering a generic list of questions about oneself to create a unique set of pages where one can create a digital representation of oneself, though the visibility of a profile depends on the site hosting it and/or the users preference. MySpace allows users to choose whether they want their profile to be visible to the public or "friends only." On the other hand, sites like Friendster and Tribe are crawled by search engines, making profiles visible to anyone regardless of whether they have an account. By default, Facebook allows a user to see the profiles of other users in the same social network, unless the owner of a profile has denied permission to those in the same social network.

After creating a profile, one is usually prompted to identify other users with whom they have a connection. The label for said connection varies from site to site, although generally they are labelled "Friends." In some cases

the connection is uni-directional: Twitter's "Follower" connection is good example of this. This means that a user can follow another user without their approval, seeing the content they produce. However, in most cases, the connection made between two users is bi-directional, meaning both users must accept the connection before they can view each other's profiles.

One of the main attractions to social network sites is the ability to communicate with other users of the site, whether that communication is public or private. Most social network sites provide some mechanism to allow these communications. Facebook, for example, allows users to post on their own profile or on a friends profile making the communication visible to all with access to see said profile although they also allow users to privately message another user.

2.3 Privacy

Social network sites have become an omnipresent technology, "tending to become invisible once widely adopted, ubiquitous and taken for granted" (Debatin, Lovejoy, Horn, & Hughes, 2009). Jones and Soltren (2005) identified flaws in Facebook that facilitated privacy breaches and data-mining. At the time, users' passwords were being sent without encryption meaning a third-party man in the middle attack could intercept and record the passwords (Jones & Soltren, 2005). Jones and Soltren (2005) also found that Facebook gathered information about users from other sources unless the user specifically opted out, which as an option was later removed forcing the data collection policy.

In the previous section, it was noted that a user could restrict who could see their profile page. This is one of the key features for privacy Facebook gives users, but this feature didn't work as intended for the first three years of its existence. Information posted on restricted profiles showed up in searches unless the user opted out of allowing searches (Jones & Soltren, 2005). This issue was only fixed after a technology blogger made the loophole public and contacted Facebook.

In September 2006, Facebook introduced the "News Feed," which tracks

and displays the online activities of a user's friends, such as uploading pictures, befriending new people, writing on someone's wall, etc. None of the aforementioned activities themselves were private actions, but the aggregated public display on the default page outraged Facebook users, who felt their privacy had been breached (Boyd, 2008). Subsequently, Facebook added privacy controls for what can be seen on the news feed and by whom.

Facebook can also be used by third parties for data mining, phishing and other malicious reasons. For example, Jagatic, Johnson, Jakobsson, and Menczer (2007) launched a phishing experiment at Indiana University on a selected group of students to gather information on students' friends using social network sites. The experiment had a 72 percent success rate within the social network whereas the control group only had a 12 percent success rate. The authors also added that other experiments in phishing on social network sites had similar results: "We must conclude that the social context of the attack leads people to overlook important clues, lowering their guard and making themselves significantly more vulnerable" (Jagatic et al., 2007).

2.3.1 Privacy Implications

The population of Facebook users studied by Gross and Acquisti (2005) is, "by large, quite oblivious, unconcerned, or just pragmatic about their personal privacy." Users generously provide personal data while using minimal limiting privacy setting are used. This visibility, variety and richness of personal information provided on Facebook, when used with the scope of the network and the public linkage to users' real identities, opens users to the risk of a variety of attacks on their online and physical persona. These risks range from identity theft to online or physical stalking, from embarrassment to blackmail. Not all of these risks are common among other social network sites, since not all social network sites use real identities or divulge as much personal details as Facebook.

Stalking

The information available on Facebook profiles can determine the likely physical location of a user. Gross and Acquisti (2005) found when researching Facebook users from an academic institution (860 profiles from Carnegie Mellon University) that students often made the physical location of their residences accessible on their profile pages as well as at least two of classes they attend. Since a students' life during university generally revolves around going to classes, it is possible for a potential stalker to know a particular student's whereabouts throughout the day. The authors also added that the research was done outside of the semester, speculating that the number of classes shown on profiles may be even higher during semester.

Re-identification

"Data re-identification typically deals with the linkage of datasets without explicit identifiers such as name and address to datasets with explicit identifiers through common attributes" (Samarati & Sweeney, 1998). An example is linking hospital discharge data to voter registration lists thus allowing sensitive medical information to be identified. A more related possible use for re-identification is using data with explicit identifiers from one social network site to identify data from another social network site. As an example, one could use uploaded images in common to link an anonymous profile from one site to a Facebook account using a real identity, thus de-anonymizing the user.

Online Harassment

Ybarra and Mitchell (2004) define online harassment as "an intentional and overt act of aggression toward another person online." Examples of this include making hurtful comments toward someone, or intentionally embarrassing another user. Finkelhor, Mitchell, and Wolak (2000) conducted a survey on internet use by youth aged 10 to 17 years of age in the United States finding that one in five were exposed to sexual solicitation, one in seventeen were harassed or threatened and 63% being stressed, embarrassed or upset while

only a fraction reported the incident. Privacy settings being used incorrectly or not being used at all can facilitate online harassment. For example, a teenager could post a message publicly for strangers to see, giving them the option and means to abuse the poster in public, which can be both hurtful and embarrassing. Abuse on the internet can lead to psychosocial trauma, emotional distress and can cause mental health consequences for teenagers (Ybarra & Mitchell, 2004). Studies are unable to show a correlation between social network sites and the increase in online harassment so it is unclear whether social network site are solely to blame or that it is the result of the use of different types of online technology as well as teenage attitudes on internet use (Ybarra & Mitchell, 2008). A common occurrence in the literature in this area is blaming the lack of parental supervision online, however, there is also considerable speculation about how teenagers perceive parents adding them to their social network online.

2.3.2 Family and Social Networks

On Facebook, individuals voluntarily disclose information, although many may not actively consider the audience to whom they are revealing information to and the variety of people included, for example the user's parents. In some cases, users may regret posting information since they did not realise before the fact that particular people (i.e. parents) are part of their social network or did not foresee peoples reactions to the revealed information. Young adults could view this as an invasion of their privacy as they feel like they must monitor what gets posted by themselves and by their friends to prevent parents from seeing. This can make young adults feel that their parent is invading conversations that they would otherwise not be privy to, particularly if the parent was to comment on the conversation. When a parent does see information about their child that they disapprove of, "boundary turbulence" (Kanter, Afifi, & Robbins, 2012) can occur between parent and child which can result in the child deleting or editing their personal information. In other cases, some young adults may not be bothered by their parent being their friend on a social network site as they feel they can use

privacy control settings to restrict what they see, although they may have little control over what their parent posts about them.

Due to the aforementioned issues surrounding parents on social network sites and the lack of empirical data on the number of parents on social network sites, it is unlikely that all young adults "friend" their parents online. At present, "individuals aged 18-25 account for the majority of Facebook users, while the most rapidly increasing demographic is individuals aged 35 and older" (Kanter et al., 2012). This could mean that parents are following a similar trend which may increase the number of family networks within social networks online.

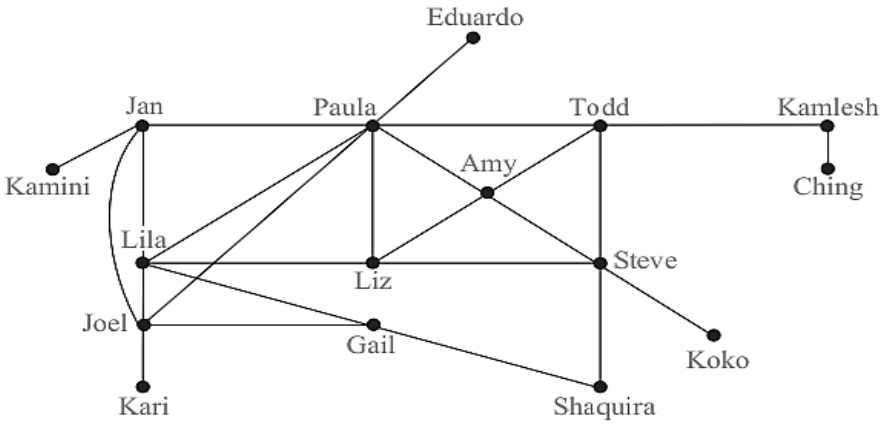
2.4 Social Network Graphs

"Graphs are discrete structures consisting of vertices and edges that connect these vertices" (Rosen, 2011). Graphs can be used to model many different problems, with many types of graphs available being used depending on the constraints of edges. These constraints vary; graphs may be directed or undirected, weighted or unweighted, cyclic or acyclic, and so forth. A directed edge is one with a direction, meaning you can only traverse that edge along the given direction, and thus an undirected edge is one without direction. A weighted edge is one with a given value, the best example of this is a road map; think of plotting a route through a country picking the best roads to take; where the vertices are towns, the edges are roads and the weights are distances. On the other hand an unweighted edges is one with no value or cost to it. A cyclic graph is one in which there are a number of vertices connected in a closed loop where an acyclic graph is one in which this can not happen. The number of nodes in a graph is known as the order of the graph, and the number edges is known as the size of the graph.

Social network graphs are a specific type of graph used to represent social structures based on the kinds of relationships between people. "In these graph models, individuals are represented by vertices; relationships between individuals are represented by edges" (Rosen, 2011). An example social network graph can be seen in Figure 2.1. The constraints, as mentioned above,

on these graphs change depend on the social dynamics being viewed. Facebook would be seen as a undirected weighted cyclic graph. It is undirected due to Facebook friendship being a mutual agreement, weighted since some Facebook friends can be added as family members and cyclic since a friend of a friend can too be a friend. Twitter, on the other hand, would be viewed as a directed unweighted cyclic graph. There are many other uses of social network graphs apart from social network sites. For example, epidemiology uses social networks to study how patterns of human contact affect the spread of diseases.

Figure 2.1: An Acquaintanceship Graph (Rosen, 2011)



There are two main algorithms available to traverse a known graph: a depth-first search or a breadth-first search. Both algorithms start from a root node. A depth-first search visits a child node before visiting a sibling node, meaning it traverses down a graph until it can not go further, then backtracks until it can move sideways to uncharted territory. A breadth-first search is often used to find the shortest path from one node to another by visiting sibling nodes before visiting child nodes. Traversing a graph becomes more difficult when the full extent of the graph is unknown. This is seen as a variant on graph traversal called graph exploration, where only the root node along with all direct edges and the nodes at the end of these edges are known. When a new node is visited, more edges and nodes can be found, revealing more of the graph.

A problem that can occur when traversing a large graph is that classic methods become too slow or fail to traverse the whole graph: in other words, these methods do not scale well. A way around this is to use heuristics to trade accuracy or precision for speed. A heuristic can be considered a short cut, using a "rule of thumb" or "educated guess" based on trial and error and data to get to a solution faster. The greedy algorithm is a good example of a heuristic algorithm; when looking for the best route, it provides a good but not optimal solution in a relatively short amount of time. It does this by picking the current best next pick regardless of whether it excludes possible future moves that are more optimal. "These algorithms effectively solve significantly larger problems than have previously been solvable using heuristic evaluation functions" (Korf, 1990). As an example of a simple heuristic: when searching through a list of names for a person's parents, there is a very high likelihood that at least one of the parents will have the same last name as the person.

2.5 Related Work

In this section, literature relating to the design and development of applications to analyse, visualise and extend social networks will be reviewed. What did they do? What are their primary results? Which challenges did they overcome? Which challenges remain? Which algorithms were developed? What analysis was done? What previous work did they rely upon? How scalable are their techniques? These are some of the questions to be answered.

2.5.1 Social Network Analysis and Mapping

Adamic and Adar (2003) designed an application which collects and analyses home pages of users to build a social network using the text on the page, the outgoing links, the incoming links and the mailing lists they are subscribed too. The text on the page provided a "semantic insight into the content of a user's page" (Adamic & Adar, 2003) which helped identify the types

of links produced by the incoming and outgoing links to other pages. The mailing lists a user was subscribed to provided a community structure the user was already a part of. The information gathered after collection and analysis was then used to visualise a user's social network. Adamic and Adar (2003) found that the application produced many false negatives, meaning a user gets matched with someone they know except there is no explicit link confirming the relationship. Another problem encountered by the authors was that not all students had personal home pages, which causes missing nodes in the mapped network, creating a lack on information. The authors noted that for the application to scale from a university setting to the entire globe would require changes to the analysis and assumptions made; this included adding more data sources such as demographic information, as well as accounting for the number of possible relationships where a user can be a fan of someone (one of thousands) or family member (one of a few). The conclusion Adamic and Adar (2003) came to was that "not only is it possible to find communities, but we can describe them in a non-obvious way."

Krebs (2002) used social network analysis to map networks of terrorist cells using data collected from news articles, search engine results and publicly released documents. The three major issues covered in their work were

- Incompleteness – the inevitability of missing nodes and links that the investigators will not uncover.
- Fuzzy boundaries – the difficulty in deciding who to include and who not to include.
- Dynamic – these networks are not static, they are always changing. Instead of looking at the presence or absence of a tie between two individuals, looking at the waxing and waning strength of a tie depending upon the time and the task at hand.

Krebs (2002) came to the conclusion that in order to successfully map a terrorist network, the agencies and/or countries involved in combating the terrorist cell need to share information and knowledge so "a more complete picture of possible danger can be drawn."

2.5.2 Recommender Systems

Given a snapshot of a social network at a given time, it is possible to predict likely interactions between users. The "link-prediction problem" as Liben-Nowell and Kleinberg (2007) put it, is "based on measures for analysing the proximity of nodes in a network." The authors show how this can be done in many different ways using already existing techniques found in graph theory and social network analysis producing varying results. They used methods based on node neighbourhoods such as common neighbours (Newman, 2001), which scores two users similarity based on the number of connections they have in common, and Jaccard's coefficient (Salton & McGill, 1983), which measures the probability that two users have a given neighbour from a randomly selected neighbour from either user. Other methods based on the ensemble of all paths were used such as the Katz coefficient (Katz, 1953), which defines a measure that directly sums the collection of paths, exponentially damped by length to count short paths more heavily. Some higher-level approaches were also used, including low-rank approximation, unseen bigrams and clustering. Liben-Nowell and Kleinberg (2007) found that "there was no single clear winner" but "there is indeed useful information contained in the network topology alone." Liben-Nowell and Kleinberg (2007) conclude by saying "there is clearly room for improvement in performance" considering the highest performing method (Katz clustering) is "correct on only about 16% of its predictions."

A similar study was done by De Meo, Ferrara, and Fiumara (2011), focusing more on the similarity between two users based on the knowledge of social ties existing among users and the analysis of activities users are involved in. They use this data to draw a local measure of similarity between the two users, then consider the network as a whole to obtain a global measure of similarity based on the Katz coefficient. Finally they combine the scores of similarity into a unique value by applying linear regression. De Meo et al. (2011) found that applying a global measure of similarity can "partially correct errors produced by each single activity" implying that a collection of techniques working together outperforms any single technique.

The authors mention future work will include applying their research to the link-prediction problem, using similarity as a measure of proximity in the network.

Kautz, Selman, and Shah (1997) designed an application, called ReferralWeb, similar to that of Adamic and Adar (2003). However, instead of focusing solely on analysing and visualising the network, a recommender system was built on top. ReferralWeb is "an interactive system for reconstructing, visualising, and searching social networks" (Kautz et al., 1997). The system is used specifically on academia for searching through papers, articles and the like to build a graph of references, allowing a user to search for academic writing by a particular person, referenced by a particular person, or referencing a particular person. Kautz et al. (1997) state that "by instantiating the larger community, the user can discover connections to people and information that would otherwise lay hidden."

2.5.3 Third Party Applications

Nazir, Raza, and Chuah (2008) developed and launched three applications using the Facebook Developer Platform in order to collect data on the usage of these applications. The applications gained a combined user base of more than eight million users, analysing the rich data procured based on geographical distribution of users, user interactions and modelling these interactions through interaction graphs. The three applications developed included a social gaming application called "Fighter's Club" in which users pick virtual fights with their Facebook friends that last from 15 to 48 hours and allow the offender and defender to request help from friends to become supporters. The other two applications developed were non-gaming applications; "Got Love" was designed to allow users to pick and display a distinct set of 'loved' friends on their profile page, "Hugged" was designed similar to that of Facebook Pokes where a user can send a virtual 'hug' to a friend (this can be done multiple times). One of the key findings is that "application dynamics can significantly affect the structure of interaction graphs, hence weakening the association between them and the underlying real-world

(friendship) relationships between users" (Nazir et al., 2008). This finding is based on the fact that users of the social gaming application would often Friend strangers in order to increase the number of supporters they can gain distorting the natural community structure. On the other hand, they found that non-gaming applications tend to exhibit strong community structures.

Facebook Friend Mapper

Facebook Friend Mapper[1] is a Google Chrome Extension designed to generate the friends list of a user using privacy settings to hide their friends list. In order for it to work, a user must have at least one mutual friend with the target user. It works by using the mutual friend to see which friends they have in common with the target user generating a partial list of the targets friends. It then uses the generated list to find users whose privacy setting allow the application to view their friends list, applying the previous method again to generate more of the targets friend list. The application continues doing this N times until a defined depth is reached, and once completed it outputs the list produced.

[1] `https://chrome.google.com/webstore/detail/facebook-friends-mapper/ikfdhlkcdllmkklmdbhfjkofjmehionn`

References

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

Boyd, D. (2008). Facebook's privacy trainwreck: Exposure, invasion, and social convergence. *Convergence: The International Journal of Research into Media Technologies*, *14*(1), 13–20.

Boyd, D., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230.

Debatin, B., Lovejoy, J. P., Horn, A.-K., & Hughes, B. N. (2009). Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication*, *15*(1), 83–108.

De Meo, P., Ferrara, E., & Fiumara, G. (2011). Finding similar users in facebook. *Social Networking and Community Behavior Modeling Qualitative and Quantitative Measurement, IGI Global*, 304–323.

Finkelhor, D., Mitchell, K. J., & Wolak, J. (2000). *Online victimization: A report on the nation's youth.* (Tech. Rep.). National Center for Missing and Exploited Children, Crimes against Children Research Center.

Gross, R., & Acquisti, A. (2005). Information revelation and privacy in online social networks. In *Proceedings of the 2005 acm workshop on privacy in the electronic society* (pp. 71–80).

Haythornthwaite, C. (2005). Social networks and internet connectivity effects. *Information, Community & Society*, *8*(2), 125–147.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*(10), 94–100.

Jones, H., & Soltren, J. H. (2005). Facebook: Threats to privacy. *Project MAC: MIT Project on Mathematics and Computing*, *1*.

Kanter, M., Afifi, T., & Robbins, S. (2012). The impact of parents friending their young adult child on facebook on perceptions of parental privacy invasions and parent–child relationship quality. *Journal of Communication*, *62*(5), 900–917.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43.

Kautz, H., Selman, B., & Shah, M. (1997). Referral web: combining social networks and collaborative filtering. *Communications of the ACM*, *40*(3), 63–65.

Korf, R. E. (1990). Real-time heuristic search. *Artificial intelligence*, *42*(2), 189–211.

Krebs, V. E. (2002). Mapping networks of terrorist cells. *Connections*, *24*(3), 43–52.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, *58*(7), 1019–1031.

Nazir, A., Raza, S., & Chuah, C.-N. (2008). Unveiling facebook: a measurement study of social network based applications. In *Proceedings of the 8th acm sigcomm conference on internet measurement* (pp. 43–56).

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, *64*(2), 025102.

Rosen, K. (2011). *Discrete mathematics and its applications* (7th ed.). McGraw-Hill.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval.

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression* (Tech. Rep.). SRI International.

Ybarra, M. L., & Mitchell, K. J. (2004). Youth engaging in online harassment: Associations with caregiver–child relationships, internet use, and personal characteristics. *Journal of adolescence*, *27*(3), 319–336.

Ybarra, M. L., & Mitchell, K. J. (2008). How risky are social networking sites? a comparison of places online where youth sexual solicitation and harassment occurs. *Pediatrics*, *121*(2), 350–357.