PRIVACY MANAGEMENT FOR ONLINE SOCIAL NETWORKS

Enkh-Amgalan Baatarjav

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2013

APPROVED:

Ram Dantu, Major Professor
Rada Mihalcea, Committee Member
Kathy Swigger, Committee Member
Philip H. Sweany, Committee Member
Barrett Bryant, Chair of the
  Department of Computer Science
  and Engineering
Coastas Tsatsoulis, Dean of the College
  of Engineering
Mark Wardell, Dean of the Toulouse
  Graduate School
One in seven people in the world use online social networking for a variety of purposes. Since social networking has rapidly become a main form of communication, holes in privacy have become apparent. It has come to the point that the whole concept of sharing information requires restructuring. No longer are online social networks simply technology available for a niche market; they are in use by all of society. Thus it is important to not forget that a sense of privacy is inherent as an evolutionary by-product of social intelligence. In any context of society, privacy needs to be a part of the system in order to help users protect themselves from others.

This dissertation attempts to address the lack of privacy management in online social networks by designing models which understand the social science behind how we form social groups and share information with each other. Social relationship strength was modeled using activity patterns, vocabulary usage, and behavioral patterns. In addition, automatic configuration for default privacy settings was proposed to help prevent new users from leaking personal information.

This dissertation aims to mobilize a new era of social networking that understands social aspects of human network, and uses that knowledge to honor users' privacy.
Copyright 2013

by

Enkh-Amgalan Baatarjav
ACKNOWLEDGEMENTS

It has been a long journey with many ups and downs. But reaching this point would not have been possible without the following people who gave me unconditional love and support. I would like to thank every one of you from the bottom of my heart.

The first and foremost, I would like to thank Dr. Ram Dantu for advising me throughout my Ph.D pursuit. He made me addicted to social networking sites, like Facebook. Ever since he introduced me to the possibility of doing research in the area, I have never left it. Thank you for that. I also would like to extend thanks to Dr. Rada F. Mihaelcea, Dr. Philip H. Sweany, and Dr. Kathy Swigger for agreeing to serve on the committee for this research.

My family is the essence of any success I have accomplished in my life. It is my great pleasure to dedicate my dissertation to my family: my grandfather, Altankhuyag; dad, Baatarjav; mom, Dondovsambuu; my brothers - Enkhbat, Lkhagvajav, Erdenebat, and Enkhtur; my sister-in-laws who I consider to be my sisters - Selenge, Ariuntuya, Delgermaa, and Oyuntulkhuur; and all of my nephews and nieces. I am very thankful for my amazing family, which was cleverly orchestrated by my visionary dad, my compassionate mom, and my honorable grandfather.

A very special person came into my life about three years ago. She opened my eyes to totally a new perspective of life; there is so much to enjoy and experience. She has become a very big part of my life and constant source of energy. I cannot thank enough for the many sleepless nights helping to reword my robotic writing to a more elegant and eligible form. Thank you again, Mallory! You rock!
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1. How Privacy is Being Handled</td>
<td>2</td>
</tr>
<tr>
<td>1.2. Motivations</td>
<td>3</td>
</tr>
<tr>
<td>1.3. Safety</td>
<td>4</td>
</tr>
<tr>
<td>1.4. Spam</td>
<td>4</td>
</tr>
<tr>
<td>1.5. Social Connections</td>
<td>5</td>
</tr>
<tr>
<td>1.6. Proposed Approaches</td>
<td>6</td>
</tr>
<tr>
<td>1.7. Dissertation Statement</td>
<td>7</td>
</tr>
<tr>
<td>1.8. Dissertation Outline</td>
<td>8</td>
</tr>
<tr>
<td>CHAPTER 2. UNDERSTANDING THE PROBLEM</td>
<td>11</td>
</tr>
<tr>
<td>2.1. Interpretation of Privacy</td>
<td>11</td>
</tr>
<tr>
<td>2.2. Real Life Social Networking</td>
<td>13</td>
</tr>
<tr>
<td>2.3. Technology and Privacy</td>
<td>15</td>
</tr>
<tr>
<td>2.4. Social Networking in the Online Context</td>
<td>17</td>
</tr>
<tr>
<td>2.5. Social Network Trends</td>
<td>21</td>
</tr>
<tr>
<td>CHAPTER 3. PRIVACY MANAGEMENT</td>
<td>25</td>
</tr>
<tr>
<td>3.1. Literature Background</td>
<td>25</td>
</tr>
<tr>
<td>3.2. Privacy Vulnerability Analysis in OSN Sites</td>
<td>26</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3.3. Privacy Protection Model</td>
<td>30</td>
</tr>
<tr>
<td>3.4. Dynamic Data Management</td>
<td>31</td>
</tr>
<tr>
<td>3.5. Privacy Management Proposal</td>
<td>33</td>
</tr>
<tr>
<td>CHAPTER 4. INTRODUCTION TO PRIVACY MANAGEMENT IN DYNAMIC DATA</td>
<td>36</td>
</tr>
<tr>
<td>4.1. Introduction</td>
<td>36</td>
</tr>
<tr>
<td>4.2. Chapter 5: Activity Correlation</td>
<td>37</td>
</tr>
<tr>
<td>4.3. Chapter 6: Vocabulary Correlation</td>
<td>38</td>
</tr>
<tr>
<td>4.4. Chapter 7: Behavioral Patterns</td>
<td>38</td>
</tr>
<tr>
<td>CHAPTER 5. DYNAMIC DATA: USER ACTIVITY CORRELATION</td>
<td>39</td>
</tr>
<tr>
<td>5.1. Introduction</td>
<td>39</td>
</tr>
<tr>
<td>5.2. Methodology</td>
<td>40</td>
</tr>
<tr>
<td>5.3. Conclusion</td>
<td>51</td>
</tr>
<tr>
<td>CHAPTER 6. DYNAMIC DATA: USER VOCABULARY CORRELATION</td>
<td>52</td>
</tr>
<tr>
<td>6.1. Main Contribution of This Chapter</td>
<td>52</td>
</tr>
<tr>
<td>6.2. Dataset</td>
<td>52</td>
</tr>
<tr>
<td>6.3. Methodology</td>
<td>53</td>
</tr>
<tr>
<td>6.4. Discussion of Results</td>
<td>58</td>
</tr>
<tr>
<td>6.5. Conclusion</td>
<td>61</td>
</tr>
<tr>
<td>CHAPTER 7. DYNAMIC DATA: BEHAVIORAL PATTERNS</td>
<td>63</td>
</tr>
<tr>
<td>7.1. Background</td>
<td>63</td>
</tr>
<tr>
<td>7.2. Methodology</td>
<td>66</td>
</tr>
<tr>
<td>7.3. Experiment Result</td>
<td>70</td>
</tr>
</tbody>
</table>
LIST OF TABLES

5.1 Friendship deduction was based on the Reply Estimator. The performance of the Reply Estimator was compared to the Twitter definition of a friend (e.g., at least two responses from a follower suggest a sign of friendship). ........................................50

6.1 Overall results of the activity correlation and the vocabulary similarity ...............60

7.1 Basic stat. ................................................................................................................ 72.

7.2 Chi Square test of independence. ...........................................................................75

7.3 Result of hour ........................................................................................................ 77

7.4 Result of the Pearson correlation analysis shows that six of seven days do support the hypothesis. Using behavioral features of the user in a given day, a set of partners can be selected to receive the status update. ........................................78

7.5 Total number of status updates were sent from the user to the partners on each day of the week ...........................................................................................................78

7.6 Result of the Pearson correlation analysis shows that 41 of 52 time windows do support the hypothesis. Using behavioral features of the user in a given day and an hour, a set of partners can be selected to receive the status update. ..............83

7.7 Total number of status updates sent from the user to the partners on each hour of the day. ...............................................................................................................84

8.1 Using the Facebook platform, application developers were able to access users' personal and social information. .................................................................86

8.2 Demographic of data revelation by gender, age, relationship, and politic preference.88

9.1 Demographic of features revelation by gender and age........................................95
9.2 Demographic of features revelation by relationship and politic preference. ...............96

9.3 Threshold matrix was built for categories of gender, age, relationship status, and political preference. .............................................................................................................................................................................97

A.1 All Twitter users have their profiles. This table shows profile features that can be accessed by the Twitter API. .................................................................................................................................115

A.2 When users update their status, the following status features are also been created.

   If a status is written to respond to someone else's status update, in_reply_to_status features are filled with appropriate information. ..................116

B.1 All Twitter users have their profiles. This table shows profile features that can be accessed by the Twitter API. .................................................................................................................................132

B.2 Vote count (VC) shows number of attributes vote to stop sampling further in the sub graph. Using the sampling algorithm sub graphs are reduced to smaller sample size (N'). ...................................................................................................................................................134

C.1 Information of 17 common interest groups on UNT social network including their subtype categories, number of members, and description. .................................................................141
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>There is a misconception that in an egocentric network, a set of random nodes circle around the center node; however, in the reality there is actually a well formed hierarchical structure.</td>
<td>14</td>
</tr>
<tr>
<td>2.2</td>
<td>Social networking sites are a union of both user-generated information and a social network of friends (the social graph).</td>
<td>18</td>
</tr>
<tr>
<td>2.3</td>
<td>Privacy management layer is located between information and social graph layers, and it assists users to control their information flow.</td>
<td>19</td>
</tr>
<tr>
<td>5.1</td>
<td>An example of possible tiers of interaction. A privacy system could use various measures to place followers into such tiers. Placing restrictions on certain tiers lessens the probability of sensitive information being misused.</td>
<td>39</td>
</tr>
<tr>
<td>5.2</td>
<td>Calculating the similarity of activity between users and followers in three different time domains.</td>
<td>41</td>
</tr>
<tr>
<td>5.3</td>
<td>Examples of activity profiles across each time domain.</td>
<td>43</td>
</tr>
<tr>
<td>5.4</td>
<td>The integral difference gives the difference in area under two curves. Difference in level of activity of Twitter users was calculated by taking the integral difference of two activity profiles. This is represented by the gray area between the two activity functions.</td>
<td>44</td>
</tr>
<tr>
<td>5.5</td>
<td>The x-axis represents followers while the y-axis represents the total-correlation. Followers are ranked by correlation.</td>
<td>45</td>
</tr>
<tr>
<td>5.6</td>
<td>This histogram demonstrates the difference between the tci ranking method and the ranking method based on the number of reply status updates. A difference of 0</td>
<td></td>
</tr>
</tbody>
</table>
means that both rankings are the same for a particular follower, while a large difference signifies very different rankings. The x-axis represents the difference measure while the y-axis represents number of users with that difference measure. It is evident that ranking methods correlate fairly well, with a large portion of users retaining a small difference measure.

5.7 An example of a user who has 357 followers with their prestige index.

5.8 The x-axis with a grey bar indicates friends (two or more directed messages sent from the user). The result of the Reply Estimator is shown in pink.

5.9 The accuracy of the Reply Estimator is shown with different friend definitions.

6.1 Distribution of post count and social graph size in the sample.

6.2 The architecture consists of three parts: activeness analysis, calculating activeness correlation, and ranking activeness correlation.

6.3 Examples of activity profiles of four users across each time domain.

6.4 The status updates similarity model consists of three main parts: preprocessing, weighting vocabulary, and calculating the vocabulary similarity between users and their followers' vector space model.

6.5 The cumulative result of the total activity correlation indexes and vocabulary similarity scores. Histograms on the x-axis and the y-axis represent the total activity correlation indexes and vocabulary similarity scores, respectively.

6.6 Applying the linear regression model to fit a line on the similarity score reveals a tendency of the relationship between the calculated similarity score and the status update count.
6.7 Side by side distribution comparison between activity correlation and vocabulary similarity. .......................................................... 61

7.1 Distribution of information is moving from a broadcast paradigm to a multi-cast paradigm. .......................................................... 64

7.2 Behavioral signature was calculated on each hour of the day. ......................... 68

7.3 Hierarchical clustering was used to split friends into two sets of receiving status updates and not receiving status updates on a time window at threshed value. 71

7.4 Similarity strength is indicated by the intensity of the color, so a darker color means more similarity between the user and that partner. .......................... 78

7.5 The heat map shows a relationship between the user and the partners which have similar behavioral patterns. .................................................. 79

7.6 The partners of the user are ordered according to their overall behavioral similarity to the user in the order of the least similar to the most similar behavioral pattern. The overall similarity score is shown in the last row in the heat map. 81

7.7 Distribution of post frequency in a 24 hour window. ..................................... 82

8.1 An example of how much personal information was revealed on the UNT social network site, based on 4,919 users. ........................................................ 87

8.2 Information revealed based on gender and age. ........................................... 89

8.3 Information revealed based on relationship status and political preference. ........ 90

9.1 Privacy management system for configuring users' privacy setting for applications. 92
9.2 Comparing the accuracy of the privacy management system using a filtered dataset of 1,746 users and an unfiltered dataset of 4,919 users. The system shows notable high error tolerance..............................................................98

9.3 The Privacy Management model was based on three components: PI, PM, and PZ. PM configured user's profile information into allowed and not allowed to access categories. ..............................................................99

9.4 Complete graph of the learned structure of B-Course. .............................................100

A.1 An example of a user's social graph on Twitter. The user is marked with dark gray color, followers are blue, and followees are orange. Edges between two identities are their relationship directions. .................................................................113

A.2 Crawler architecture consists of four parts: manager, recovery, crawler, and data storage. ..........................................................................................................................122

B.1 Example of a social network graph. In the graph there are six sub graphs marked by purple, and each one has a source node marked by blue node. There can be multiple sub-subgraphs Ai with in sub graphB. .................................................130

C.1 Basic architecture of GRS, which consists of three major components including profile feature extraction, classification engine, and nal recommendation. ........140

C.2 (a) Gender ratio of each group. (b) Age distribution ranges of 15 to 19, 20 to 24, 25 to 29, and 30 to 36. (c) political preference distribution of the following very liberal (VL), liberal (Li), moderate (M), conservative (C), very conservative (VC), apathetic (A), and libertarian (Ln). ....................................................... 143
CHAPTER 1

INTRODUCTION

Imagine if there was a way that you could call every single person on your phone contact list simultaneously. *What would you talk about?*

This question is posed every day on online social networks, and there are both positive and negative aspects about it. For one thing, it is an efficient outlet for others to reach out to people with whom they have a relationship, but at the same time, people are not always comfortable with sharing details about their daily lives with everyone they know.

Online social networking (OSN) is an environment that provides a single destination for anything to be discussed, posted, and shared. It can be described as a wide open field of personal information, family, friends, colleagues, strangers, opportunities, disasters about to happen, uncensored feelings, relationships, high school reunions, photos from vacations, cats, missed opportunities, crushes, love, marriage, divorce, anger, grassroots movements, food, and alcohol fueled parties. Online social networking, in essence, tries to capture every aspect of a person’s social life. Ultimately it is a tool that enriches our every day experiences with an interesting spectrum of ideas from people we know.

However, as a limited tool, online social networks also potentially allow for the collection of private information, discomfort of users, stress, and the possibility of monetary and physical harm. These issues cause users to want to know how and where their information is shared when they post something, before they feel ready to share it. If an automatic mechanism is built to be both aware of user privacy concerns and also understand its users’ relationships, it could ensure that messages are delivered to the correct audience. This would allow users to finally embrace online social networks as a viable medium of communication.

Current models in online social networking sites lack understanding of how human interactions and relationships work and therefore were not developed with user privacy in mind. If these models do not address how and to whom information is being posted, online
social networking will never be a primary choice of communication and will continue to be the root of many societal problems.

1.1. How Privacy is Being Handled

The power to spread information to anyone with very little cost is the basis of social networking. Existing online social networking (OSN) sites provide an information broadcasting service. Users write short sentences answering the question ”What’s on your mind?” on Facebook, ”Share what’s new” on Google+, or ”Compose new Tweet...” on Twitter; the answers for these questions are called status updates. When the users finish writing their status updates, they are broadcasted to their friends (social graph) and possibly to the public as well. Depending on the users’ privacy setting and social network sites’ policy, status updates are distributed differently. How existing social networking sites handle privacy of these users’ status updates is discussed below.

1.1.1. Facebook

In Facebook, users can broadcast their status updates to five different sets of their friends. A set can be Public\(^1\) (everyone on the Internet), Friends, Friends except Acquaintances, Only Me, or custom (manually created lists of friends). The default privacy setting is configured to Friends, such that all of the friends in the social graph can see those updates. In order to be privacy conscious while posting updates, users have to create custom lists of friends, such as family, close friends, acquaintances, etc. Facebook users tend to have around 229 friends on average, so assigning each friend into one or multiple lists can become burdensome very quickly. Thus, the feature is not effectively used.

\(^1\)Status updates posted as public can be indexed and searched by a search engine. For example, Openbook was created to harvest public status updates on Facebook and functioned between May 2010 and May 2012 [13]. Without careful consideration of privacy settings, the Public setting can lead to unintended exposure of private status updates to the whole Internet.
1.1.2. *Google+*

One of the most recent social network sites is Google+, started in 2011. The main design principle is based on friend circles, where everyone is assigned into one or multiple friend circles. Status updates can be shared with specific circles, Public, or Extended circles (circles of circles or friends of friends). The main problem of sharing information with circles is that it is almost as demanding as using the Facebook list feature, despite it having a more user-friendly UI.

1.1.3. *Twitter*

Contrary to G+ and Facebook, Twitter uses a binary approach in privacy settings for status updates. Status updates can be either public (everyone on the Internet can see them) or private (only approved followers are allowed to see them). The default privacy setting is on public.

1.2. **Motivations**

An ideal online social network is one where people are able to share and communicate easily with each other. This means that the direction of future internet technology needs to depart from closed and independent communities, and instead move towards an open and connected society. Openness creates collaboration, diversity of opinions, and freedom of information. Online social networking (OSN) is one of many assets of life that has been moving our culture closer towards an open community. Judging by the rate of the subscription to OSN sites today, the size of this community will only get bigger and and stronger in the future. As of September 2009, the size of Facebook community (300 million) passed the U.S. population, only taking 5 years and 6 months to do so. After that, it only needed five more months to reach 400 million users. Today there are more than *one billion* users on Facebook [20]. OSN sites have revolutionized the way we communicate, the way we share

---

As a side note, both public updates on Facebook and Twitter were available through Google real-time search, which was disabled shortly after it was launched on Dec 7, 2009 [124]
information, and the way we do business. Since OSN sites are so relevant to our daily lives, this is something which should be embraced.

1.3. Safety

In real life, we do not share our daily activities with every person we know because it is not safe. Instead, we share it with people whom we trust and with whom we have common history. In other words, it is a selective process that we can control. However, the same cannot be said for OSN environments since the technology is beyond an average user’s control. OSN sites promote a very open community: anyone can befriend anyone else [144]. Therefore it is difficult for any user to comprehend the extent and speed of the information spread. And unfortunately, if users trust OSN sites to protect their privacy, they are more willing to reveal their information [60]. In spite of these challenges, users should be aware of the risks and have much more privacy control for their posted information. Without effectively addressing these safety problems, users can become subject to online predators, cyberbullying, spam, and ultimately will have problems trusting the system.

1.3.1. Predators

Studies from the Journal of Adolescent Health have shown that 82% of sex predators have found minor’s likes and dislikes on social networking sites, 65% have found home and school information, and 26% have even been able to locate their victims’ exact time and location through OSN sites [104].

1.3.2. Cyberbullying

The OSN environment also inadvertently creates a haven for bullying, as 81% of youth believe cyberbullying to be much easier than bullying in person [16]. A 2011 Consumer Reports statistic shows that in the past year, at least one million children have been harassed or threatened through cyberbullying on Facebook [14].

1.4. Spam

If we look at status updates from the receivers’ perspective, it can be overwhelming to read through all of the status updates they receive in one day. In Facebook, these status
updates come as part of news feed. In addition to status updates, it contains updates of friends’ activities on Facebook, such as using Facebook applications, making changes in profile and in social circles, etc. Average Facebook users have 229 friends [20] and receive about 73 or more status updates on average.

Another problem of getting a large number of status updates is that relevant updates are usually overlooked. Having a privacy management layer will help to broadcast posts to a specific audience that is interested in the topic. Irrelevant status updates have shown to have a negative effect on recipients’ quality of the experience from the news feed.

Without clever filtering, a large volume of information presented to users via the news feed is overwhelming. As OSN sites become popular with more people, this already large volume of information will increase sharply, and this has a negative correlation of increasing irrelevant information. A similar analogy to this problem is spam e-mail, which a 2009 Microsoft Security report showed 97% of all emails sent over the Internet to be unwanted [142]. Without a good filter for spam/junk e-mails, e-mail would not have been a main medium of communication. With the same token, the content generated on social networking sites is becoming a nuisance and has the same pitfalls as spam e-mail does.

1.5. Social Connections

Being able to accept and be friends with just about anybody online creates the risk of users not being able to trust the system. Sharing information is an intimate social activity that involves knowing with whom to share what kind of information. Humans have an innate ability to channel specific information to the right audience in real life. However, socializing online blurs the normally clear boundary of social groups and lumps everyone into one group. This is an invasive approach to organizing social groups, and users do not know their social graph (who their friends are / how many they have). Therefore, it reduces users’ trust in the system, causing them to be reluctant about posting.

OSN sites cannot exist as a main medium of communication if users feel uneasiness when sharing. When communication takes place between multiple partners, it is much more complex and harder to adopt the system. Other technologies which have become main
mediums of communication are a good source of inspiration, such as the telephone and cell phone. Since there is little fear of privacy intrusion, they were more easily adopted.

If social networking keeps going as it is presently without addressing privacy concerns, it could miss out on opportunities in performing smoothly as a main form of communication. Fully enabled, it could tap into the psychology of genuine social networking and generate revenues for OSN providers.

1.6. Proposed Approaches

This dissertation concentrates on developing social metrics in order to tap into information about social relationships, social interactions, and social tie-strength. By understanding this information, a privacy management system can be organically developed to protect privacy. An important consideration when approaching this problem is that social metrics should be adoptable on different OSN environments because the available attributes are different in those environments.

The model proposed in this dissertation can be used in a number of applications in the online social networking environment, such as in status updates, location updates, and photo updates. It can also reduce the information overload currently being experienced in OSN sites. Since it appears that none of existing OSN sites have such privacy measures, it could be an important improvement for providers such as Facebook [3], Google+ [7], MySpace [8], bebo [9], hi5 [10], and Twitter [4].

Another important problem addressed in this dissertation is protecting the personal information of new users, since OSN sites tend to use the opt in model. The opt in privacy model creates vulnerabilities by causing new users to reveal their profile information unknowingly. For any new software, there is a learning curve for any user, which is a period of time for the users to understand the system and learn to navigate it fluidly. Since some OSN sites make the configuration of privacy settings very hard to find and understand, it can be challenging for even experienced users.
1.7. Dissertation Statement

The motivations in this dissertation are concentrated on two components. The first component is to change how information (status updates) is broadcasted on (OSN) sites, such that users’ privacy is protected and users are able to control who may see their updates; the second component is to design an automatic configuration of the default privacy settings for profile information.

Once these two forms of private information are protected properly, spam will be reduced, trust will be initiated in users, and minors can avoid contact from predators and cyberbullies. If these changes can be made effectively, online social networks will easily become main modes of communication to the benefit and convenience of society.

1.7.1. Main Contributions

All of the following contributions are related to the context of online social networks:

- Developed a universal matrices to quantify social relationship strength
- Designed a model that used timestamps in the status updates to compute tie strength of social relationships [29]
- Created a model that used status updates to rank social relationships among friends [31]
- Demonstrated that users’ activity patterns extracted from timestamps give a better prediction of social relationship strength, rather than vocabulary usage in the status updates [31]
- Analyzed the usage of special keywords in status updates to build a mode for deciding the right audience for a given status update
- Proposed the default privacy configurations for new users using the knowledge of established users [33, 32]
- Presented software architecture for developing applications on Twitter network
- Showed an application of Kolmogorov complexity in sampling large scale social networks
Developed a group recommendation system for Facebook [34]

1.8. Dissertation Outline

The dissertation explores a number of ways to improve existing online social networks by using a social context of human behavior, such that interaction on the online social network would be more natural. The following list shows the roadmap of this dissertation:

**Chapter 1**: This chapter addresses the problem definition faced currently in online social networks and explores the motivation behind the need for improvement. It shows the overview of ideas presented in the dissertation and includes the thesis: How does one attempt to manage privacy in online social networks?

**Chapter 2**: Before delving into proposing solutions for this problem, the background behind the issues is explored through various lenses in other fields, such as anthropology, social science, and psychology. It includes the philosophical interpretation of privacy, the structure of human social networks and its formation. It discusses personal boundaries, how real life social networking works, and how online social networks have created a new social environment. Stats of online social networks are shown in order to demonstrate how they have become a primary form of communication, as well as how an online social network is structured and how it is broken down into different parts.

**Chapter 3**: This chapter presents and overview of the issues involved in privacy management, and how technology has affected the concept of privacy. It provides a literature survey of relevant papers on privacy management. It proposes a way to capture human nature in social interaction in order to improve privacy. It also gives a brief overview of the proposed solutions, discussed in more detail in other chapters.

**Chapter 4**: This chapter is an umbrella for chapters 5, 6, and 7, covering privacy management in dynamic data.

**Chapter 5**: This chapter discusses how to estimate the social affiliation strength between a user and friends. One obvious way to estimate social strength is to look at the frequency of message exchanges between a user and friends. A new way to
estimate social strength was discovered using behavioral patterns of the user and the friends. The result of this study concluded that users and friends who have similar activity patterns are socially close friends.

Chapter 6: New ways to estimate social strength between users and their friends continue to be explored in this chapter. Natural language processing (NLP) was applied. The hypothesis was that socially close friends spend a large amount of time together since socially close friends share similar interests and topics. To test this hypothesis, the usage of same vocabulary words by both user and friends were analyzed, so if friends have high social ties, they are apt to use the same vocabulary words. The result concluded that usage of vocabulary could be a good indicator to estimate the strength of a social tie; however, it also produced high negative correlations. Vocabulary patterns and activity patterns were compared to estimate the social tie, and result of the comparison was that activity patterns were a better estimation.

Chapter 7: This chapter investigated using behavioral patterns when selecting a set of friends as an audience for a given status update. For instance, status updates that are posted during work hours are more relevant to colleagues then friends. The hypothesis was tested on two different time domains: day and week, and result of the study showed promising results that in a given time window, a set of friends can be selected based on the similarity of activity patterns in that time interval. However, a concrete conclusion was not able to be drawn due to a small data set.

Chapter 8: This chapter was an introduction for the next chapter covering privacy management in static data by analyzing how much users reveal information according to their demographic.

Chapter 9: In this chapter, a method was proposed to protect the privacy of static data on social network. Using a general tendency of specific profile information, privacy management for Facebook is modeled to configure new users’ privacy settings. The experiment was conducted on Facebook data, and the result of the study
achieved 75%. Additional improvements were made using BBN, and the result boosted to 90%.

**Chapter 10:** Discussions of the proposed method are included, and the limitations of the work are addressed. Possible future improvements to fundamentals in online social networks are proposed in how content is filtered. Overall interpretation of my dissertation work and limitation of the work are included in this chapter.

**Chapter 11:** The dissertation is concluded with a summary of the contributions and the future of online social networks.

**Appendix A:** This is where a user guide for developing a Twitter application that specifically collects data is shown. The architecture of the software design is included.

**Appendix B:** A possible approach to designing a crawler using Kolmogorov Complexity to sample social networks is presented.

**Appendix C:** One of the issues in social networking is that there are often too many groups for a user to select, and finding an appropriate one is difficult. A recommendation system is proposed based on the similarity between a user’s profile and a group’s profile.
CHAPTER 2

UNDERSTANDING THE PROBLEM

In order to tackle the difficulty involved in protecting user privacy in OSN sites, two main components need to be first explained. Understanding real life social networks and how they are formed are very important when examining the popularity and seeming innocuousness of online social networks. The other component involves both defining why privacy matters and how it functions in real life.

2.1. Interpretation of Privacy

Privacy can be explained by the evolutionary theory that humans developed a systematic mechanism to protect ourselves from our environment. This ideology is shown through Herbert Spencer’s concept of the survival of the fittest [127].

In the context of human evolution, the survival of the fittest can imply both physical fitness and intellectual fitness. Intellectual fitness defines the process of making use of available information for our own benefit; in the sense of survival of the fittest, however, it makes us vulnerable to others within our own species. This is because anyone in the same species likewise has nearly the same intellectual capability to process available information and act accordingly.

As a strategy to survive in our competitive environment, we want to accentuate our strength and divert others’ attention away from our weaknesses. Thus, in the modern world, information has become a valuable commodity that should only be distributed or shared carefully.

2.1.1. Personal Boundaries

In a stable society, having privacy be respected by other people is a natural right that exists in the form of personal boundaries. These boundaries allow partners in conversation to have a notion of which topics are acceptable to discuss, and which are not. By continually receiving feedback during conversations, people acquire skills which help them to recognize
personal boundaries. Children are an example of what happens when personal boundaries and privacy are ignored, as they often say impolite things because they have not yet learned to avoid certain topics depending on the situation.

If the privacy boundary is disregarded, there have shown to be specific patterns for emotions we feel. If a stranger asks a personal question regarding physical health, the first reaction for most people would be to wonder \textit{why} this stranger wants to know that information, and then the body will display signals of irritation or displeasure. If the stranger is not able to read that body language and react to it accordingly, an awkward moment occurs and the conversation is discontinued. However, the same question when asked by family members or close friends will often be received much differently, sometimes even resulting in a positive reaction.

In general, positive reactions promote more social interaction between members of a social graph, while too many negative reactions will eventually surpass a certain limit. Breaching this limit can cause an individual to withdraw from the interactions until balance is reached, which Schwartz defines as the \textit{threshold} of privacy.

In order to create a stable social system, it is necessary to have a well established guarantee of privacy protection. Without that assurance, individuals participating in social interaction feel discouraged from involving themselves in any social activity as they have had their threshold breached. The nature of a social establishment provides not only shelter and a comfortable living environment, but also the assurance of privacy. Therefore, privacy has to be designed as a part of the physical structure from conception.

To illustrate this concept, a typical office room can serve as an example since it is designed to have four walls and a door. This creates an environment which ensures the privacy of the owner since he or she can choose to make himself/herself visible by others by opening the door. Personal items are also arranged on the desk or in the drawers such a way that secures private information.

Most people maintain an invisible bubble of personal space to keep themselves feeling comfortable in social situations. This is defined by Edward T. Hall as \textit{proxemics}, which
describes the intimacy level of relationships between people as being correlated with physical distance [75]. Basically, proxemics is nonverbal communication which shows how people, even in everyday life, keep a comfortable physical distance away from others.

This physical distance varies depending on the social distance of people. Hall’s study categorizes the amount of physical space between people into four groups, arranged from least to greatest:

- Intimate distance
- Personal distance
- Social distance
- Public distance

The intimacy level and cultural differences between people also factor into establishing personal space. People who have intimate relationships maintain a closer distance and use physical contact. In personal distance, which is the physical distance reserved between family members and close friends, a slightly farther apart distance is used. Social distance and public distance are kept for interactions between acquaintances and public audiences, respectively.

Proxemics indicate that there is a socially acceptable physical distance between people. So in the context of the privacy, it can be interpreted there is a certain respect of privacy when people interact with each other in real life. How much of that privacy can be shared depends on the intimacy between the interacting individuals.

2.2. Real Life Social Networking

Because privacy and personal boundaries touch on multiple aspects of human life, literature from other fields of science need to be examined.

2.2.1. Anthropological Perspective

From the anthropological perspective, human beings have a natural tendency to form social networks consisting of hierarchical groups [145]. Figure 2.1 shows how human social
network structure is often misconceived. This phenomenon has been observed in both human and non-human primate systems [82, 103].

The study of social distance was first conducted by Hediger, who was a Swiss biologist and is considered to be the father of zoo biology. He observed that animals of the same species tend to maintain specific distances away from each other, which he called personal and social distance. There was also a certain distance maintained when animals were around a different species, called flight and critical distance [80].

Hall’s study of proxemics, or "personal space” also concluded similar behavior where humans’ hierarchical groups are arranged into social subgroups based on their intimacy and the frequency of their interactions [75]. Another interesting phenomenon in group size formation is that the size of these groups follows a geometric progression with a scaling factor of 3, such as 3, 9, 27 [145]. This means that each human hierarchy grows in size in a constant factor as they become socially distant.

2.2.2. Social Brain Hypothesis

Another argument of why humans form hierarchical group structures to maintain their social graph may be explained by the social brain hypothesis, where a cognitive constraint exists of 150 individuals with whom social relationships can be maintained at the same time.
Social relationships in this context imply that humans not only know the names of the friends, but also are keeping track of all social interactions with those friends.

Being able to effectively maintain social relationships requires a high computational capacity of the brain in any given species, which a number of studies supported by showing a high correlation between a social group size and relative neocortex volume [58, 56, 121].

In order to simplify the complexity of maintaining a social network and to optimize its efficiency, a hierarchical group structure is desirable. The most intimate and closest group is the smallest, but individuals in this group are the most trusted family members and close friends with whom the most confidential details are shared. Most individuals spend the majority of their social time with this group [59].

As social groups move down in hierarchy, the social capital decreases while the number of people in the groups increases. These kinds of social hierarchy structures can be observed in nature [82], such as in the grooming clique size of hamadryas [95] and the frequency of their interactions within the social graph [57, 59].

2.3. Technology and Privacy

Technological advances have always been an important driving force behind changes in the way society perceives privacy [126]. Two phenomena in particular have shared the property of causing information to be visible to the masses via technology: the printing press and the Internet.

These technologies function as mediums of communication and were made possible by technological advances throughout history. The first medium of the two was the rotary printing press, which was invented in 1843. It was steam powered to make mass quantities of copies which were then distributed in great quantity to the public. As the printing press became a main mode of sensationalized news, it began to breach the privacy of popular figures (otherwise known as the origin of yellow journalism and paparazzi).

In order to restore balance the system, laws and regulations were made to give protection to the victims. The first publication that addressed privacy by assuring individuals’ right to privacy was published by Warren and Brandeis in 1890 [141].
A second wave of technology was the Internet, where basically any information became instantly available to reach the mass public with one click of a button. There are now many social networking sites that are available to connect people and allow them to share personal information. Not only has the number of social networking sites been growing, but also the range of services has been diversifying as well.

The significance of social media is starting to be realized and is valued to be a significant part of the Internet’s future. For instance, the number of patents related to social media have sharply increased in number from less than 25 in 2003 to more than 1200 in 2010 [2].

Social media has become a communication tool that connects family members, friends, colleagues, and strangers, allowing everyone to share information among different groups. Social media so far has forced itself into our everyday lives, but in order for it to be successfully integrated into our society it has to provide privacy protection tools to support the integrity of its users. The control of these privacy tools has to be more organic and designed with the perspective of the users’ in mind [23].

Managing user privacy is not necessarily a simple task as people have different expectations of privacy, depending on the situation, audience, and personality. There is no one glove which fits for everyone, so any efficient method will require some of the functions to be automated. This will help users to not be overwhelmed by the process of privacy configuration, though it is difficult to find a balance between overwhelming and underwhelming [130].

It may seem from the perspective of social network providers that giving control of privacy to users is not a good move in terms of profitability. It is with the hope of this research, however, that understanding how social interaction works in society will prove to benefit both users and providers since ultimately users will be more willing to share information. It will actually nurture and encourage communication among family and friends because better privacy control gives assurance that the information being posted will not be mishandled.
The next ten years of social media will most likely be concentrated on adjusting social media to fit better with the expectations of society. Anatomically speaking, modern humans have been on the earth for about 200,000 years, having evolved and perfected the way social groups interact. From this standpoint it would not be wise for OSN providers to attempt to change how interactions should work overnight.

2.4. Social Networking in the Online Context

Now that a real life context for social networking has been addressed and how technology is a new driving force was explained, the implications in online social networking (OSN) sites can now be looked at in further detail. Much more than in real life situations, OSN sites have created a new environment to extensively connect people with their social graph.

Online social networking has so far been a collection of chaotic information and connections among individuals. These networks do not have good management of the information nor are they able to control the flow of information among individuals. To begin resolving the root of the problem on social networks, understanding the essential parts of a social network is critical.

In essence, social networks are a marriage of two fundamental parts: information and the social graph, which is shown in Figure 2.3. Information is defined as being any content which is provided by users of the social network. Based on the update frequency, this information can be divided into two types. Information that does not change or is not updated frequently is called static information, and it includes profile data. The other type of information is something users often update and is called dynamic information; this includes things like status updates and GPS data.

Static information identifies who a person is overall and it does not change very often because it contains core aspects of a person. This information includes profile features, such as name, age, gender, likes, hometown, religious belief, political positions, etc. and is usually filled out when users create an account.
Dynamic information represents the behavior of a person over time in the form of status updates, GPS location, photos, and activity patterns. Characteristics of this type of information are that it is often personal, it is relevant to small subsets of people, and that it is short-lived.

The content of both types of information can be anything in a wide range of categories, ranging from highly sensitive information to non-sensitive. Sensitive information should be seen only by a socially intimate group of individuals, while non-sensitive information can be shared with everyone in the social graph.

The second part of a social network is the social graph, or the network of individuals and their connections to other people in their lives, such as strangers, acquaintances, childhood friends, family, schoolmates, close friends, and many other types of people, and it is maintained by the individuals themselves. The dynamic of any given social graph always changes overtime, but the size of the graph tends to grow so much that at a certain point it
is too complex to maintain [81]. In response to the complexity of maintaining a social graph, self-similar groups with hierarchical structures are formed in the network [76].

In an egocentric network, the center node forms its surrounding nodes into self-similar groups based on social relevance, shown in Figure 2.1. There is a clear pattern of frequency interactions in an egocentric network between a center node and subsequent within. If the pattern of interactions within a user’s social graph is studied, there is a psychological behavior in the frequency of interactions between the user and the user’s friends, such that there is a hierarchical structure of groups, shown in Figure 2.1. A number of studies conducted on social networks support the phenomenon that the frequency of interaction follows the power-law distribution, which is that the frequency of interactions are very high with only a small set of friends, and it declines sharply with the rest of the friends.

In order to successfully merge these two critical parts (information and the social graph), there needs to be third aspect of social networks, called privacy management, which

![Figure 2.3. Privacy management layer is located between information and social graph layers, and it assists users to control their information flow.](image-url)
is a layer sandwiched between the information layer and the social graph layer. The main goals of this privacy management system are to create an organic way to handle privacy in the social graph and to automate the process by using learning behaviors.

It is human nature to manage our social network in hierarchical groups from intimate to distant relationships. Because the social network consists of multiple layers of social groups, information being shared needs to flow into the correct groups. Personal information that is relevant to the most intimate group should only be seen by the individuals in that group. Assuring that the information flows to the right social group will actually help social networking to be a natural medium of human communication and to remove the escapism of posting status updates of our life on social networks. This dissertation aims to explore different ways to find social intimacy through examining individuals’ behavior and afterwards integrate this data into a privacy management of a given social network.

2.4.1. **Numbers**

So much data about human interactions has been generated through OSN sites because of their extreme popularity. OSN sites can be found for business connections [12], dating [6], photo sharing [11], video sharing [15], music broadcasting [5], microblogging [4], mainstream networks [3], bookmarking [1], etc. Three of the most popular OSN sites are Facebook, Google+, and Twitter; as of February 2013, Facebook is the largest social networking site with more than 1 billion monthly active users [20], and in December 2012, Google+ and Twitter had 343 million and 288 million active monthly users, respectively [105]. Google+ is a relatively new social networking compared to the other two, but its user base has been growing steadily since its debut in June 2011.

OSN sites have revolutionized way we communicate and share information. In 2007, 5,000 status updates\(^1\) were being posted on Twitter in a day. The average number of status updates per day has increased dramatically in the following years [17][18][19]:

**2008:** 300,000

\(^1\)On Twitter, status updates are called tweets.
2009: 35 million
2010: 50 million
2011: 140 million
2012: 340 million

Today, there are about 600 status updates being posted every second. This number will increase as more and more people adopt this mode of communication. In fact, based on socialnomics, 96% of the Millennials,\(^2\) which is estimated to be around 80 million people, have joined a social network site[116]. A study conducted between October 2008 and February 2009 by Inside Facebook [125] shows that the fastest growing demographic is women over the age of 55. During the time of study, this demographic’s usage on Twitter grew by 175.3%. This indicates that OSN sites are being embraced by not only teenagers and young adults, but also by older generations.

2.5. Social Network Trends

Social networking is becoming an exciting field of study, especially since it has been under the influence of many different trends. In this section, the main trends of social networking are addressed, including events from the past, present, and future [30].

2.5.1. Early Trends of Social Networking

Online social media has been undergoing changes as readily as the underlying infrastructure of the internet changes. The concept of OSN sites started as early as 1995\(^3\). However, the content and social network of these sites were much different than how they are today, since they instead consisted mostly of static information and a few socially interactive features. Tripod.com[128], theGlobe.com [97] and GeoCities [24] were pioneers in building these online communities by providing limited web-hosting space in which users could create their own content. The GeoCities network was organized into a pre-defined set of neighborhoods, forcing users to select a neighborhood to which their web pages were the

\(^2\)A generation that was born in between 1977 and 1995. It is often also referred to as Gen Y, Generation Why, Echo Boomers, Generation Next, and Gen I (Generation Internet) [134].

\(^3\)This is referred to as the first generation of online social networks.
most relevant. Even though the early states of the social networking sites offered features that promoted social interactions, such as chat, message boards, and classified sections, they lacked an efficient social network structure.

The next trend involving social networks arrived by the end of the 1990’s. The main features of this generation were the inclusion of personal profiles, the integration of social connections, and functions for browsing/searching friends on a network. The user interface of OSN sites became much more user-friendly, so users no longer needed to understand how to build a website design from scratch - instead, they were able to fill in standardized forms to create personal profiles. The users’ engagement in OSN sites improved greatly from earlier trends since communication with friends was easily accomplished through the usage of bulletin boards, e-mail services and online instant messaging. Ultimately, this created a community of users who could share similar interests.

An early developer of this secondary trend was SixDegrees [37], which, after being launched in 1997, reached one million users by 1998; unfortunately it closed its services in 2000 [42]. There were a number OSN sites which immediately followed suit of SixDegrees, but they were targeted toward different demographic groups. AsianAvenue (1999), BlackPlanet (1999), Migente (2000), and Cyworld (2001) were created for specific ethnicities, Ryze (2001) and LinkedIn (2003) for professional and business networks, and MySpace for entertainment and music. Now there is an even wider range of OSN sites targeting different interest groups [42].

2.5.2. Current Trends of Social Networking

The current trend of social networks started in 2004, with the emphasis being on both building a social network platform that allowed for a variety of applications to access it, and renewing content to improve users’ engagement. Facebook changed how social networks are perceived: no longer just a web application that connects users’ profiles within the network, it has become a platform that can be integrated into web applications, allowing for marketing via the users’ social graph.
Most of the major social networking sites now offer their platforms for developers, such as Hi5, Orkut, Friendster, Facebook, and Twitter. Having an open platform has made it possible for the integration of many different interfaces to social networks. Not only can Facebook and Twitter be accessed through the web interface, but also through desktop applications, smart phone applications, and SMS. Sharing photos, location, and status updates has become much easier, which has encouraged some smartphone manufacturers to start building phones that have dedicated a Facebook button on their keyboard for easier access [135].

It has become more and more evident that OSN sites are putting a lot of emphasis on the dynamic content that is constantly updated by users. One way of doing that is through manipulating aspects such as the Timeline on Twitter, News Feed on Facebook, or Stream on Google+. Dynamic content is much more valuable than static content because it is both updated constantly and actively engages the users on the site. This trend was first pioneered by Twitter, whose main function revolved around broadcasting short and current information to mass audiences. The idea was later inherited by most of the other OSN sites. It is now the main feature of many OSN sites.

2.5.3. Future Trend of Social Networking

The future trend of social networking will emphasize more ways to share dynamic content. The total amount of dynamic content generated on OSN sites is already large, but it will become even larger as more users join these online social network sites. Although dynamic content is an attractive feature, many of the items shown there are simply not relevant enough for any user to want to spend a lot of time reading. Additionally, relevant updates from closer friends are easily overlooked since there is such a large volume of information to sort through.

How information is distributed on social networks also has to be improved in order to protect users’ privacy. Considering the vast amount of content being generated on these social networks, it is evident that users are willing to share their information to families, friends, and strangers. However, a deeper problem here is that older ideas involving the
privacy of shared information do not support this natural tendency, which inhibits users from posting. Google+, on the other hand, has capitalized on this idea by creating friend circles, so that users share their information to only selected friend circles, ultimately giving the control to the users.

Even though it has taken 13 years for people to become acquainted with sharing information on OSN sites, there is still some skepticism because of its lack of sufficient privacy and security features. The proposed privacy management layer will hopefully help users to feel more comfortable sharing, as it will control the spread of information. To make the model more flexible when working in different environments, such as different social networking sites, the model needs to be able to use different types of data: static, dynamic, and the social graph. The following chapters discuss issues in privacy management and explore aspects of static data and dynamic data collected from different OSN sites in relation to privacy management.
CHAPTER 3

PRIVACY MANAGEMENT

Online social networking has been revolutionizing the way people communicate, as its popularity has been growing sharply in all generations. Unfortunately, however, this amazing growth and popularity gain is beginning to prove that it is too good to be true. After all of the initial hype cooled off, the reality of danger began to creep in. After a number of incidences [104, 16, 14] in which privacy intrusions occurred at the hands of online social networking sites, people are now more reluctant to share too much personal information. Online social networking sites can actually capture the essence of our lives: from infants to elders, from waking up in the morning to going to bed at night, and from family to strangers. The more people realize this, the less they will be willing to use social networks to their full advantage.

This dissertation utilizes dynamic and static information found on online social networks to create a privacy management model with the intent to address some of these problems. In this chapter, current literature on privacy issues is discussed. Chapter 4 is an introductory chapter for dynamic privacy management and an umbrella chapter for Chapter 5, 6, and 7. The approach to static information is introduced in Chapter 8, which is followed by a proposal for static privacy management in Chapter 9.

3.1. Literature Background

Online social networking (OSN) sites have become an integral method of communication when connecting to family and friends. They also facilitate the sharing and gathering of information with strangers, who may or may not have similar interests. Practically, however, this is not always the case. Thus, effectively managing privacy on social networking sites has started gaining more and more attention from individuals in many fields of study; in particular, computer science, information science, and sociology. Because the number of publications related to this topic has grown considerably, it is now somewhat challenging to
do a thorough literature survey. Although this survey has been narrowed down to publications from 2007 and 2012, it is not, by any means, exhaustive. The current research can be divided into three main domains: privacy vulnerability analysis, privacy protection models, and dynamic privacy management.

3.2. Privacy Vulnerability Analysis in OSN Sites

Average OSN users can be easily manipulated if they believe that OSN sites are entirely secure, and therefore are more willing to post personal information [60]. A study done on the usage of privacy settings shows that a majority of Facebook users do not change their default privacy settings even though they are able to limit the visibility of their profile information from strangers [73]. Additionally, having a large amount of personal information available in easily harvestable environments like Facebook, MySpace, and Orkut can lead to social phishing attacks [87]. Phishers can impersonate the friends of victims or use personal and social information in other harmful ways.

Accordingly, a number of surveys have been done to find out how much personal information is revealed by average users. In 2007, Acquisti and Gross (Carnegie Mellon University) reported their research about privacy concerns arising from Facebook social networking [21]. Their survey of 40 questions relating to Facebook privacy was taken by 506 respondents. They analyzed the behavior of the survey-takers on Facebook based on before and after learning how information can be revealed on Facebook. They also pointed out the misconception of members’ profile visibility in their network based on the survey.

Social networking has become increasingly popular among teenagers, but, unfortunately, they are also notorious for easily becoming victims of privacy invasion [41] [146]. This stems from their relative lack of concern (or ignorance) about the importance of maintaining privacy, which has proven to lead to both physical and online attacks. To address this issue, Susan B. Barnes proposed three approaches to solve this problem: social, technical, and legal [36]. Danah Boyd’s study presented four properties: persistence, search-ability, exact copy-ability, and invisible audiences that OSN sites have, but conventional face-to-face interaction
does not [41]. These properties have been changing the way people interact, especially for young people.

In a research study conducted by Krishnamurthy and Wills [93] on 3,851 randomly selected MySpace users, 1,600-1,700 Facebook users\(^1\), and other OSN sites, the main privacy leakage sources came from default privacy settings, users’ utilization of privacy settings, and third-party applications. To analyze privacy leakage, user’s profile information is grouped into five categories from the least private to the most private: thumbnail, greater profile, list of friends, user generated content, and comments. Utilization of privacy setting reveals that 79% (3046 randomly selected) of MySpace users retained their default privacy setting, 80% on Bebo, and 80% on Facebook\(^2\) as well. Finally, monitoring HTTP requests and response showed that the majority of the most commonly used OSN sites exchanged HTTP messages with third party advertisers and data aggregators, such as doubleclick.net, 2mdn.net, advertising.com, atdmt.com, googlesyndication.com, etc.

Strater’s and Richter’s study [129] quantified college students’ disclosure and privacy behavior on Facebook with the intent to develop a better future privacy system. Participants completed demographic surveys and a personal inventory (NEO-FFI). Each participant evaluated two other participants’ profiles. The following is the highlight of the study: 77% of the participants had a publicly accessible profile. Even though the participants were aware of both the privacy concerns associated with Facebook and how to manually configure their own privacy settings, many participants still maintained the default setting, which is that the profile would be publicly accessible. The participants exhibited “all or none” approaches to disclosing their personal information.

- 77% of participants had a publicly accessible profile.
- 33% had a private setting for their profile.

\(^1\)Users were sampled from 20 different U.S. regional networks and 20 different non-U.S. regional networks.

\(^2\)In a global level, 80% of users kept their default privacy setting. However, people who are on U.S. Facebook networks were more privacy conscious than Non-U.S. networks; 29% of users on U.S. Facebook networks changed their default setting, while only 10% of Non-U.S. Facebook users did.
• Even though participants were aware of privacy concerns associated with Facebook and the configuration of privacy settings, many participants maintained a default setting (profile was publicly accessible).
• 25% did not disclose any information in personal fields while 67% filled in all personal fields.
• Amount of personal information revealed significantly affected the evaluation of participants. Information disclosure about interests, favorites, and "about me" were positively correlated with the rating of trustworthiness.
• Participants exhibited "all or none" approaches to disclosing their personal information.
• 67% indicated pictures as the most communicative feature of a profile because they were a self-presentation of users.
• By providing a clear view of the social network, friend proximity, and availability of profile features, visualization tools may improve usability of privacy configuration.

Dwyer et al. [61] showed ways in which OSN site users control their online privacy and also how to generally improve the privacy management system. 222 members of the New Jersey Institute of Technology (NJIT) participated in this study, with 107 of them being Facebook users and 115 subjects being MySpace users. 18.9% of participants suffered a privacy incident, and less than half of them took counter measures to protect their privacy by changing their privacy settings. Their conclusion was that "privacy must be conceptualized as a quality of an online space, rather than as a collection of access settings to be managed by individual members." Integrating privacy as an integral part of OSN sites, they recommended the following suggestions: evaluating the privacy level of each component, providing privacy feedback, and publishing privacy norms.

Schrammel et al. [123] studied 850 online survey takers’ information disclosure behavior and the correlation between information disclosure and demographic background. The study included different online communities: business networks (e.g. LinkedIn and Xing), social networks (e.g. Facebook and MySpace), content and media sharing networks (e.g.
Flickr and YouTube), and social news and bookmarking sites (e.g. del.icio.us and Digg). The information disclosure behavior on different networks was analyzed based on nine common attributes among the different networks. Some of the highlights were that students were more willing to disclose their information on business, social, and social news and bookmarking networks, but not on content and media sharing networks, indicating that revealing information was directly related to the level of trust users had in the network. Males disclosed more information to their friends than females on social networks, including real name, nickname, picture of user, date of birth, network of friends, email, physical address, phone number instant messaging contact, and link to the website of the user. Some of the interesting findings were:

- Students were more willing to disclose their information on business, social, and social news and bookmarking networks, but not on content and media sharing networks.
- Revealing information was directly related to trust in the network.
- On social networks, the more skilled and experienced users were, the more willing they were to disclose their information to strangers.
- Males disclosed more information to their friends than females on social networks.
- In general, users’ occupation had more influence on information disclosure behavior than gender or age.

Based on a dataset of 35,000 users, Lindamood et al. [100] argued that OSN users are vulnerable against attacks which use inference. They carried out an experiment based on 35,000 Facebook users, modifying the Naïve Bayes algorithm to classify a large amount of OSN data. The algorithm took node traits and link structure as its inputs to predict a various user’s political preference as either liberal or conservative. The experiment result was studied in four different cases: without removing any predictive trait and links, removing the 10 most predictive links and none of the traits, removing the 10 most predictive traits and none of the links, and removing the 10 most predictive traits and links. Their algorithm
performed better than the traditional Naïve Bayes algorithm and Links only algorithm. It also performed better than the Details only algorithm in most cases.

Another inference study using 66,766 personal profiles from the Livejournal network was conducted by He et al. from the University of California in Los Angeles [78]. They used a Bayesian network approach to learn social relations on OSN sites, and they called it the Bayesian inference. The Bayesian inference outperformed the traditional Naïve inference. Their study of influence strength and society openness revealed both that personal attributes are highly correlated to the strength of the relationship between a user and the user’s social graph, and that an effective way to protect privacy disclosure is to selectively hide friendship relations or friends’ attributes. A unique idea in this study was that they define the influence of the social graph by the number of hops between a user and the user’s friends.

3.3. Privacy Protection Model

Users’ data, such as behavior, interests, and demographics is valuable for personalized systems and web applications. The data, when harvested, can be used to adapt to each user’s personalized needs. Even though personalized systems are beneficial in providing relevant contents, targeted emails, and e-commerce, Internet users continue to express significant concerns about their privacy[132]. In the general context of the Internet, a lot of research and projects have been done to address the issue of protecting privacy. For example, the Platform for Privacy Preferences (P3P) project was created by W3C for a privacy standard [51]. They proposed that user agents should be integrated into browsers in order to check for the compatibility of users’ privacy preferences and their website privacy policies. There have been several improvements made in the following research [113] [133] [112] [114].

A number of cryptographic approaches address the vulnerability of OSN users’ privacy: Lucas and Borisov propose placing an encryption/decryption structure on social networks so that only users in the same social graph are able to decrypt information[102]. A variation of the cryptographic approach was also proposed in NOYB (None Of Your Business) by Guha et al.[74]. Anderson and colleagues[26] propose a client-server architecture using cryptography for social networking to ensure users’ privacy.
In an extension of the previous work, *Inferring Private Information Using Social Network Data* [100] [79], Heatherly et al. proposed mitigating techniques to deflect inference-based attacks. The main argument for this study was that Facebook privacy configurations do not guarantee users’ privacy, especially since a third party can predict undisclosed information of users based on just the friends of any user.

3.4. Dynamic Data Management

Advancement in Internet technology has revolutionized the ways online social networks are being used today. One prime example was Web 2.0, which contributed in making social networks a much more engaging and dynamic environment. It is no surprise that the popularity of Web 2.0 coincided with beginning of Facebook in 2004. The main change in social networks since 2004 has been that content of information has been moving toward more dynamic data.

On Facebook, dynamic content is generated under name of newsfeeds, where users answer the question "What’s Happening?"; meanwhile users on Twitter answer the question "What are you doing?". The content emphasis of Facebook has been shifting from profile information to newsfeeds. Meanwhile, the sole service of Twitter is around status updates. Twitter has been surprisingly successful for a company which does only one specific service, but this is supported since social networking today is purported to revolve around dynamic content in order to keep users engaged [42].

Privacy control for dynamic content is much different than general privacy control. Once the general privacy control is configured for static data, it is apt to stay the same for a long period of time because the information it protects and audience for that information does not change overnight. On the other hand, privacy control for dynamic data has to be extremely flexible and able to be configured depending on the content. This is because the content has the potential to be either sensitive, so that it should be shared among close friends, or insensitive, so that it can be shared with everyone.

Just as privacy control for static information started to be integrated after the public outcry surrounding profile privacy concerns, privacy concerns regarding dynamic information
will become the next big issues in using social networks. Google+, one of the latest social networks, is based on a framework of social grouping to broadcast status updates to a specific group. This is a praiseworthy step toward understanding human nature in sharing content on the social networks, as it implements a system which makes use of the knowledge in order to make a better social networking experience[91]. Unfortunately, however, it is still difficult to use because it requires user input.

Privacy control for social networks can take advantage of the structure of a network that can be clustered into groups, proposed by Jones and O’Neill [89]. Their study showed that users employed six criteria of social circles: tie strength, temporal episodes, geographical location, functional roles, and organizational boundaries to cluster. Knowing these criteria made it possible to build an automated system using a network clustering algorithm SCAN, and one privacy configuration was applied to all of contacts in a same group. The proposed system’s top performance was 76.1% of contacts correctly configured to right privacy settings.

The current development of mobile devices which can provide exact users’ exact GPS coordinates is pushing the envelope of privacy concerns. One area of service which requires a well designed privacy management system is a location based service. Designing a system for broadcasting location information has to address a number of design problems: how to balance granularity of the location information, how to effectively configure privacy settings, how to define the willingness of the users to reveal their locations, and how to design configuration profiles for different social groups [120, 27].

Privacy control for managing access and denial can be very complicated due to a lot of factors that can be used to define personal preference, such as current location, time of the day and week (month and year), what a person is doing, and whom that person is with currently. Capturing the context of a situation when developing a privacy model has proven to be a very complex task. Current research findings were mostly on understanding the problems through survey [28] and manual configuration of privacy policy level based on physical location [45, 106].
Fang and LeFevre proposed a machine learning approach to privacy configuration and called it a *privacy wizard*. The mechanic of their approach was that after users were asked to configure their privacy settings for a given set of their friends, the *privacy wizard* would be able to configure the rest of the friends who had similar features and network associations and were also configured with similar privacy settings. The essence of their idea is that there are implicit rules that can be learned using machine learning techniques, so that privacy settings can be managed. The study was conducted on 45 Facebook users, and the result of the study was that the *privacy wizard* was able to achieve an accuracy of 90%. [63]

The content of social networking has been becoming more and more diverse as technology advances. The amount of information it can capture is virtually endless, and it can make that information available to anyone immediately. This creates a never-before seen social environment. It also is the source of many privacy concerns, and it is rightfully so. Fortunately, research communities have begun addressing these issues, alongside social network services that are also starting to take the right steps in integrating privacy management systems. However there is still much to be discovered in this topic, especially since dynamic information, a newer aspect of online social networking, poses even more new challenges in protecting privacy. As of the time of this dissertation, only a limited amount of research has addressed privacy concerns in dynamic data in particular. Thus, it is hoped that this dissertation can provide some further insight by delving into privacy management of both dynamic and static information.

3.5. Privacy Management Proposal

One major area that current online social networking models have failed to understand is the physiology of social interactions. This dissertation explores ways to improve social networking by capturing the human nature of social network structure, as well as the psychology of privacy in privacy management models that can be adapted to fit each individual. Two privacy management models for both dynamic and static data on social networks are proposed.
The first asset on social networks is dynamic data, and protecting this data requires much more maneuvering. The content of dynamic data, such as status updates, can contain a wide range of topics. While some topics are sensitive and should only be seen by an intimate group of friends, the others can be public. It is a much more difficult problem to select the right group of friends to whom the message should be shown. That was when it was realized that configuring privacy in dynamic data has to depend on the context of the message and the social relationships between user and friends.

Ways to define social relationships between a user and his or her friends in a given context was explored, using a correlation of their social behavior similarities. It was discovered that people who are socially close do share a high similarity in their behaviors, especially in terms of vocabulary usage, and also their activity patterns on an online social network.

The second asset of OSN sites is static data, and many social networking sites do offer privacy control for static data such as profile information; a problem, however, is that people who are new to the system do not utilize their privacy settings, and by the time they realize such features, it is too late because all of their information was configured to be public by default and had already seen by the unwanted parties. To protect against such scenarios, ways to utilize crowd sourcing were explored in order to configure default privacy settings for new users, which is, in a sense, transferring wisdom from experienced users to new users.

In the study, a 75% accuracy was achieved in configuring settings of default privacy. Further improvements were made using Bayesian Belief Network, and the result of this study on a small sample of data reached 90% accuracy.

Existing OSN sites have concentrated on the technological aspects of a social network, such as gathering information and connections. However, these sites severely lack an understanding of human nature in social relationships, especially in terms of defining the strength of social ties, formation of social groups, and sharing behavior within the social graph. This dissertation studies the social science of relationships between individuals, the formation of social groups into a hierarchical structure, the philosophical understanding of privacy, the evolution of privacy, technological advances, and negative and positive consequences of a
social network. The nature of this work entails an interdisciplinary study of social science and its integration with new communication technology, such that the experience of social networking can be improved to have human-like behavior in social interactions.
CHAPTER 4

INTRODUCTION TO PRIVACY MANAGEMENT IN DYNAMIC DATA

One type of information available on social networking is dynamic information, which consists of status updates and location information. Managing privacy of this type of information is much more challenging due to the characteristics of the data. In the context of social networking, the content of the information tends to be changed and updated frequently, and it often contains very short and informal language. The audience for dynamic data can be anyone within a wide arrange of social groups, from family and close friends, to strangers. The main challenges are defining social groups, determining the strength of social connections, and configuring the adaptability of a privacy management system to mesh with different types of social networking sites. In the following chapters 5, 6, and 7, ways to address these challenges are explored in order to build social metrics.

4.1. Introduction

The recent rise of massive online social networks have spurred scholars, political activists, and advertisers to investigate the nature of human social interactions and the propagation of information within such networks. Providers such as Facebook, Google+, and Twitter allow users to build a significantly large social network by increasing the number of individuals with whom users can connect. These established social connections are typically labeled as "friends", "followers", or other similar terms. However, as the list of connections expands, it becomes increasingly difficult for users to manage their interactions and preserve the privacy/exclusivity of their shared information. These issues arise because the network embodied by these connections does not represent the actual interactions and relationships that a user maintains with others [67]. Social networks usually weigh each connection equally whereas real social relationships fall under many different classes based on interests, intimacy, work, etc. This leads to privacy problems as certain information may be accessible to the wrong users [50][52][90]. If a client or server side application was created that automatically excludes certain users based on their relevance, it could greatly reduce privacy concerns.
Furthermore, the sheer number of connections forces most users to focus on only a small number of people that reciprocate attention and are deemed most important. For example, Golder et al.’s study of Facebook showed that in an online social network setting, the word "friend" has a very loose meaning. "Friend" can mean a person whom a user went to school with together, met one time at party, or met in an online chat room. Users also only poke and message a small number of people, despite having a much larger number of declared "friends" [71]. A casual search of recent calls made through any mobile phone usually reveals that only a small percentage of the names stored in the phone are frequently contacted by the user. A study of social interactions within Twitter reveals that the driver of usage is a sparse and hidden network of connections underlying the "declared" set of friends and followers. In the following three chapters, these social ties will be analyzed and classified according to the strength of a given relationship.

**Definition** The strength of a social tie is the combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services that characterize the tie [72].

The links within a user’s general social graph can be split into weak ties and strong ties [72]. Strong ties include people such as family, close friends, and people who can affect one’s emotional health [122]. Loose acquaintances can be categorized as weak ties. Worded differently, the strength of a social tie can be expressed by the following features: communication reciprocity, social overlapping, intimacy, frequency of communication, structure of social graph, and demographic [40]. If the strength of a social tie can be extracted, online social network services can provide users with enhanced functionality to meet demands of privacy and management. The next three chapters explore different ways to address this issue.

4.2. Chapter 5: Activity Correlation

In Chapter 5, social tie strength was defined through examining activity correlations. The study was inspired by the work of Berscheid et. al [38] who determined that people who have a close relationship do indeed have similar activity patterns. In the study, activity
patterns were extracted and correlated between users and their friends. The experiment of the study supported that even individuals’ activity patterns on OSN sites can be a good predictor for the strength of social ties. In addition to the activity correlation study, a way to model probability was explored by determining how likely a user’s friends are to reply to a given status update, based on that user’s popularity (prestige index).

4.3. Chapter 6: Vocabulary Correlation

After studying activity correlations in order to define the strength of social ties, additional ways to find social metrics were explored in Chapter 5. In the chapter, behavioral patterns were examined involving vocabulary usage among users. Socially close individuals tend to spend more time together, sharing common activities, interests, and discussion topics, so vocabulary used in their status updates should also reflect that similarity. Interestingly, the hypothesis was supported by the evidence of the experiment. However, when comparing the experiment results between activity correlation and vocabulary correlation for determining the strength of social ties, activity correlation proved to be a better indicator for defining the closeness of social relationships.

4.4. Chapter 7: Behavioral Patterns

In Chapter 5, a more granular approach was investigated when selecting a set of friends with whom to share a specific status update. Users’ behavior was defined based on their usage of certain keywords in their status updates. For a sample with a large volume of exchanged messages, the results of this study were useful in determining how relevant status updates were to a user’s friends. However, the results were inconclusive if the dataset did not contain very active users.
CHAPTER 5

DYNAMIC DATA: USER ACTIVITY CORRELATION

5.1. Introduction

In the face of information overload, how can Twitter (and, by extension, other OSN sites) better manage its users’ privacy? As mentioned previously, the problem lies with the assumption that all "friends" within a social network are equal. This assumption precludes the fact that users have different levels of interaction with these "friends". In reality, Twitter users can place their followers and followees within certain tiers of interaction, shown in Fig. 5.1. Close friends and family can be placed in Tier 1 while distant acquaintances or celebrities fall in the higher tiers. Adding functionality that helps users place followers into such tiers could enhance Twitter's user experience by increasing manageability and privacy. This chapter investigates various measures, which gauge the strength of a particular social tie, and then incorporate them into a system that will improve tractability and privacy management for Twitter users.

![Diagram of possible tiers of interaction]

**Figure 5.1.** An example of possible tiers of interaction. A privacy system could use various measures to place followers into such tiers. Placing restrictions on certain tiers lessens the probability of sensitive information being misused.
The system will consist of two tools that deal specifically with manageability: the Activity Correlation and Reply Estimator. Activity Correlation determines the most relevant followers. Relevance is based on whether the followers are active during the times when the users are active. Followers who are online when the user sends a message are more likely to respond at that time. More importantly, similar activity profiles could correlate to a stronger social tie. This study examines these relationships and uses the knowledge to be able to build an Activity Correlation.

The Reply Estimator refines this process by computing the willingness of a Twitter member to reply to a message from another Twitter member. This willingness value, or reply estimation, is based on the popularity of the receiver and sender (the prestige index), the activeness of the receiver, and the level of interaction (direct or indirect) between the sender and receiver. By knowing the reply estimation of the receiver, the sender can then decide whether or not to proceed with the message.

5.2. Methodology

Figure 5.2 shows the methodology of estimating the activity of users and followers based on hourly, weekly, and monthly time domains. In each time domain, the similarity of user’s and follower’s activity patterns is calculated and ranked from the most similar activity patterns to the least. After calculating the results for each time domain, the data was combined into one ranking called the Activeness Ranking, showing the overall similarity of activity patterns.

The following steps were employed:

1. Build an Activity Correlation meter to extract a relevant set of followers. This meter will help to place followers into the various tiers of interaction.
2. Build a Reply Estimator to further refine the set by gauging the likelihood of response from a follower. It is assumed that a greater reply estimation suggests a closer friend.
3. Incorporate these tools into a privacy setting that restricts information to a targeted set of followers that are deemed safe.
5.2.1. Dataset

To find social closeness between users, a dataset was compiled from numerous Twitter feeds using the Twitter API (available for third-party developers). A PHP crawler was developed in order to communicate with the Twitter server. The volume of data handled by the crawler was fairly large, causing several challenges along the way, including memory leakage problems in PHP and program crashes due to mismanaged memory and unexpected data feeds from Twitter. Significant work was required to recover from the crashes.

In this study, 3,652,148 status updates, 7 users, and 2,305 followers were analyzed. 2,045 and 1,128 of the status updates were reply messages sent from user to follower and from follower to user, respectively. The average active account was 276 days old, which was calculated by subtracting the time stamp of the first status update date from time stamp of last status update date.

The collection process proceeded in two stages. First, Twitter users were selected with the following constraints: users must have had between 100 and 200 followers, and they must have had more than 50 status updates. Based on these constraints, a total of 5,995 Twitter users were selected for which both user profile data and status updates were collected. The second stage entailed collecting user profile and status update data for all of
the followers. Collecting status updates for the followers occupied most of the time spent; overall 8,000 follower status updates were collected.

5.2.2. Activity Correlation Meter and Experiment

To build the Activity Correlation meter, points of differentiation were found among followers. On one level, there were two types of followers: followers that had engaged in multiple directed exchanges with the user, and followers who had not. In trying to predict social closeness, sifting through followers who had not had direct exchanges with the user became quite challenging. The methodology used to build the Activity Correlation meter analyzed and quantified correlations between user activity and follower activity. First, activity profiles based on hourly, daily, and monthly timescales were constructed for both users and followers. Then correlations were calculated between the activity profiles of the users and those of the followers. Finally, all correlation data was totaled and interpreted for meaningful trends that could point to stronger social ties, the idea being that the more correlated the activity profiles, the more likely that the user and the follower shared common interests.

The Activity Correlation meter sought to shrink the set of followers to something more tractable in terms of manageability and privacy. The process used activity profiles to estimate the strength of a social tie. The activity profile characterized user activity on Twitter through both the time and frequency of the status update. It was hypothesized that highly correlated activity profiles would suggest a stronger social tie between a user and a follower.

To construct the activity profiles, a number of status updates were plotted over different time periods. Fig. 5.3 shows three examples of activity profiles, one for each time domain. It is evident that there were significant trends within each time domain. Therefore, a comparison of user profiles should yield interesting results.

The first step in the comparison fit the activity profile to a curve. For the interpolation method, the data points were fitted to a polynomial function $p(x)$ of degree six:

$$p(x) = \sum_{i=1}^{n} p_i x^{n+1-i},$$

(1)
Figure 5.3. Examples of activity profiles across each time domain.

It is believed that the model worked well for the comparison because there were relatively few data points, and the curve was flexible enough to accommodate trends. A higher degree was found to oscillate too much, hiding relevant trends. Next, the integral difference of the functions were calculated, representing the user and follower profiles:

\[
\text{id}_{u,f} = \int_{1}^{N} (p_u(x) - p_f(x)) dx,
\]

The integral difference enabled the ability to quantify how much the activity profiles between user and follower deviated. The number of intervals in each time domain was denoted by N. The day, week, and year domains had the intervals 24, 7, 12 respectively. Polynomial function \( p_u(x) \) represented the users activity profile data points, and \( p_f(x) \) represented the followers activity profile data points.

Figure 5.4 displayed the two functions and the integral difference represented as the area between the two curves. The integral difference, however, only took into account the amount of deviation between the two curves, neglecting their actual correlations. Therefore it was necessary to measure correlations between the two curves. The correlation coefficient was calculated for each time domain, \( cc_{u,f} \). The correlation coefficient showed the strength and linear relationship of the user’s and follower’s activeness levels, denoted by covariance \( \text{cov}(u, f) \):

\[
cc_{u,f} = \frac{\text{cov}(u, f)}{\sigma_u \sigma_f} = \frac{1}{N - 1} \sum_{i=1}^{N} \left( \frac{u_i - \bar{u}}{\sigma_u} \right) \left( \frac{f_i - \bar{f}}{\sigma_f} \right),
\]
The integral difference gives the difference in area under two curves. Difference in level of activity of Twitter users was calculated by taking the integral difference of two activity profiles. This is represented by the gray area between the two activity functions.

In this equation, the sigmas denoted the standard deviations of the user’s and follower’s activeness levels. The correlation coefficient will always fall between values -1 and 1. The result should be interpreted as follows: $cc_{u,f} = 1$ high correlation between $u$ and $f$; $cc_{u,f} = 0$ no correlation between $u$ and $f$; and $cc_{u,f} = -1$ high inverse correlation between $u$ and $f$.

Together, the integral difference and the correlation coefficient gave a complete relation between user and follower activity profiles. These two measures were combined into a single quantity called the correlation index:

$$ci_{u,f} = cc_{u,f}(1 - \frac{id_{u,f}}{\max(id_{u,f})})$$

The integral difference was normalized by the maximum $\text{abs}(id)$ of all the followers to give a consistent result between -1 and 1. A correlation index was calculated for each time domain. However, for the purposes of this analysis, it was desirable to have a total correlation that quantified the correlations over all time domains:

$$tci_{u,f} = ci_{u,f}^{\text{hour}} + ci_{u,f}^{\text{week}} + ci_{u,f}^{\text{month}}$$

The $tci$ was normalized between -1 and 1. $tci = 1$ suggested a strong relation between the activity profiles of user and its follower. Using $tci$, followers were ranked based on the similarity between users’ and followers’ activity profiles. An example is shown in Figure 5.5 for a particular user and the corresponding followers.
Figure 5.5. The x-axis represents followers while the y-axis represents the total-correlation. Followers are ranked by correlation.

In order to understand the usefulness of the correlation index, this ranking was compared with a control ranking. This control ranking ordered the followers based on the amount of responses made between the follower and user. The more responses there were, the stronger the social tie. Therefore, if the total correlation index is an effective measure, there should be a strong similarity between the two ranking mechanisms. The comparison was achieved by taking the difference of both rankings. If two followers were ranked similarly, their differences should be fairly low. The differences from all users were combined and presented in the histogram, shown in Figure 5.6. The figure shows that a significant portion of the followers were ranked similarly by both ranking mechanisms (the difference is close to zero), suggesting that the activeness correlation could be useful in determining the strength of a social tie. However, the variance is fairly large, hurting the accuracy. The ranking system will hopefully be improved for better accuracy.
Figure 5.6. This histogram demonstrates the difference between the tci ranking method and the ranking method based on the number of reply status updates. A difference of 0 means that both rankings are the same for a particular follower, while a large difference signifies very different rankings. The x-axis represents the difference measure while the y-axis represents number of users with that difference measure. It is evident that ranking methods correlate fairly well, with a large portion of users retaining a small difference measure.

5.2.3. Reply Estimator and Experiment

To compute a reply estimate, data was collected from Twitter containing information about users. Specific attributes included the following:

1. Number of followers and followees (those the user follows) declared by the user.
2. UserID, screen name, and the utc offset.
3. Status update count for each user (both direct and indirect).
4. Active time since registration.

Definition A member receives many directed ties, but initiates few relations. The prestige index quantifies the popularity of a particular member based on the number of followers and number of followees: $PrestigeIndex = \frac{\text{number of followers}}{\text{number of followees}}$. By analyzing the relationships between factors such as prestige index and status update counts (for both directed and
indirect updates), a possible model for building a Twitter Reply Estimator is suggested. Central to the analysis will be the added label of “friend” on particular followers based on the number of directed responses received by those followers.

Definition A friend is defined as any follower that has received two or more directed messages from the user. This is similar to the first set of followers defined for the Activity Correlation meter [86]. In addition it is assumed that a friendship exists if follower follows a followee, and a followee follows a follower.

The final metric is the reply estimation, which is a percentage calculated from the prestige index, response type (the frequency of responses over period of activity), and relationship.

The Twitter Reply Estimator would take Activity Correlations a step further by gauging the likelihood of a response, thus adding dimensionality to the ranking system in the previous section. Twitter users usually have a lot more followees than followers. With the average number of followers numbering in the hundreds, a constant stream of Twitter feeds would become overwhelming. The simplest filter would be to respond to only those followees who have directed a message directly to the user. The Reply Estimator would sift through a user’s set of followees and determine which one would be most likely to respond based on certain statistics.

As mentioned previously, the primary statistic to be incorporated into the Reply Estimator is the prestige index, so the first task would be to better understand the demographics of the Twitter population in terms of this prestige index. Figure 5.7 displays a distribution of users according to their prestige indexes, with much of the mass lying between prestige indexes 0 and 1. This confirms the intuition that most users follow more people than they are being followed by, thus yielding a ratio of less than one. Of course there are users that lie above one, suggesting that they are more popular. For example, there is only one user who has prestige index greater than 1.

It was discussed previously that users will only focus their attention on other members who reciprocate that attention. Based on this idea, an assumption was made as to which
followers are most likely be the user’s friend. Specifically it was assumed that any follower to whom the user directed two or more posts would be considered a friend, while all other followers would be simply considered followers. A measure of the reply estimation was then used to compare the set of friends with the more general set of followers. The reply estimation was calculated as a function of relative prestige index (How much does a user and followers prestige index differ), response frequency, and reciprocity. This probability measure was based on the following assumptions:

1. Let a relative prestige index be \( \frac{\text{Follower Prestige Index}}{\text{User Prestige Index}} \). The higher this ratio, the lower the probability of response from the follower.

2. Probability of response is directly related to the Update Frequency - the amount that a follower updates within a given period of time.

3. A reciprocal tie between two Twitter users suggests a greater probability of response than a directed (one-way) tie.

The results can be seen in Fig. 5.8. Again, we see that the response estimation is higher among followers that have a history of responding to the user. This suggests that response estimation is a useful measure in targeting these followers. Next, the performance of Reply
Figure 5.8. The x-axis with a grey bar indicates friends (two or more directed messages sent from the user). The result of the Reply Estimator is shown in pink.

Estimator was tested against different friend definitions by changing the number of direct messages sent from the user, shown in Figure 5.9. Although there is a slight accuracy decline for tightly defined rules for friend definition, the overall performance is fairly high, between 97% and 93%. The saddle point $x = 16$ indicates that the number of direct messages sent from user to the follower are usually less than 16. The accuracy of Reply Estimator can be defined by the Equation (6):

\[
\lim_{F_{D \rightarrow 56}} (\text{ResponseEstimator}) = 93%,
\]

Also, Figure 5.9 reveals that the number of direct messages sent from the user to the follower are usually less than 16, which demonstrates that Twitter is mainly used for broadcasting information.

The Reply Estimator was further analyzed with friend definition 2 (two or more messages sent from user to follower) in Table 5.1. Only 5% of followers fit this definition. Although some users had many followers, they contacted only a small percentage of them.
The accuracy of the Reply Estimator is shown with different friend definitions.

The nature of the Twitter network is that one can follow anyone, which may explain this phenomena. TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

Table 5.1. Friendship deduction was based on the Reply Estimator. The performance of the Reply Estimator was compared to the Twitter definition of a friend (e.g., at least two responses from a follower suggest a sign of friendship).

<table>
<thead>
<tr>
<th></th>
<th>Followers</th>
<th>Friends</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>103</td>
<td>22</td>
<td>17</td>
<td>80</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>User 2</td>
<td>220</td>
<td>2</td>
<td>1</td>
<td>217</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>User 3</td>
<td>410</td>
<td>1</td>
<td>0</td>
<td>409</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>User 4</td>
<td>317</td>
<td>5</td>
<td>4</td>
<td>311</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>User 5</td>
<td>174</td>
<td>2</td>
<td>2</td>
<td>167</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>User 6</td>
<td>357</td>
<td>35</td>
<td>30</td>
<td>297</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>User 7</td>
<td>724</td>
<td>46</td>
<td>44</td>
<td>656</td>
<td>22</td>
<td>2</td>
</tr>
</tbody>
</table>
5.3. Conclusion

The methods proposed by both the response estimator and the activity correlation meter rely on calculating the strength of the social tie between a user and follower. Given a definition for "friend" (such as a follower with two or more directed responses), the response estimation and activity levels correlate highly with followers that are considered friends. These two tools form the basis of a system that could enhance the privacy of Twitter feeds by restricting certain information to a relevant set of followers.

Much of the challenge in undertaking this analysis was during the data collection process. Future work will deal with the design of the proposed tools based on the results of this chapter. However, this suggests that several unanswered questions would prove crucial in using these studies when creating a privacy system. Further tests are required to adequately understand the significance of the activity correlation measure. The ranking method suggests a relationship between activity and response frequency, but the variance in the results indicates that other validation methods would be useful.

As of the time of this research, several users have setup the configuration such that when somebody registers to follow them, they automatically follow the follower. This artificially balloons up the list of followers. The end goal is to identify real friends who will respond compared to recreational followers. Another goal is to give users better control over broadcasting their status-updates. Mechanisms proposed would most likely be automatic. Default settings could limit the extent of a broadcast to those followers that the system deems safe (based on various metrics). This would ensure a default privacy setting that can handle sensitive information in a more intelligent manner.
CHAPTER 6

DYNAMIC DATA: USER VOCABULARY CORRELATION

In this chapter, building the social matrices with restrictive constraints is continued, such that it is adaptable on most social networking environments. The previous chapter explored how to use an activity correlation to define social strength. The result of the study showed that users who have strong social ties also have a strong activity correlation with their friends. This chapter extends this work in order to analyze the usage of vocabulary as a quantifier for the strength of a social tie. The experiment concludes by comparing the results of the activity correlation and the vocabulary usage correlation in order to define the strength of the social tie.

6.1. Main Contribution of This Chapter

Existing social networking sites lack sufficient privacy management, as they are unable to define social relevance among friends/followers and broadcast status updates. Thus, this experiment explores just how to find a socially relevant set of a user’s followers. However, the question that remains is this: Which attributes help to define social relevance?

In order to create an effective model that blends seamlessly into the framework of real social networking sites, the attributes used in the model need to be available to function in many different types of environments. An example of these common attributes are status updates and the timestamp of status updates, since most social networking sites offer these features. Two universal attributes considered in this particular study are activity correlation and vocabulary usage similarity.

6.2. Dataset

The dataset used in this study was provided by Microsoft Research. It contained 1.3 million conversations gathered from the Twitter Public API from July 1, 2009 to August 27, 2009, including 477,045 unique conversations and 296,486 Twitter users. Conversations were built by following the ⟨in_reply_to_status_id⟩ tag from the API.
Due to limited resources, the stratified sampling approach was applied using ten different strata. The following steps were taken in order to sample a smaller portion of the data: first, the total number of users were grouped into a strata based on the number of conversation partners they had. Second, the strata was ordered in descending order, with the two 10% extremes from the top and the bottom being removed since an assumption was made that the two extremes of the strata were noise. Third, 40% of each strata was randomly sampled for the study, since it is thought that this amount can properly represent the data set. The distribution information of the sample is shown in Figure 6.1. Figure 1(a) indicates the distribution size of users’ social graph, and Figure 1(b) indicates the distribution of post count in a histogram with 50 bins.

![Figure 6.1. Distribution of post count and social graph size in the sample.](image)

6.3. Methodology

Although some attributes that define social relevance are rather obscure, a careful analysis of just how social interactions are formed has revealed hidden patterns that can be integrated into a model. These attributes were discovered by analyzing the activity and vocabulary usage between users and their followers. First, followers were ranked based on social relevance, which is defined by the number of status updates sent from users to followers. This ranking is referred to as the base rank. Second, the activity patterns for users and their
followers were calculated. For vocabulary usage, the same steps were applied by ranking followers based on the base rank and calculating the vocabulary similarity between users and their followers’ vocabulary words. The relationship is then calculated by quantifying the activity correlation and their base ranks, and the vocabulary usage similarity and their base ranks.

6.3.1. Activity Pattern Analysis

The nature of the study in this chapter is to identify additional social metric to define social strength of users' relationships. In this part the analysis for activity patterns, the same model was applied as in Chapter 5.2.2; however, the time domain here is reduced to two: day and week domains. Only a brief overview of the methodology was given in this section, and detailed a explanation of the methodology can be found from Chapter 5.2.2. The summary of the architecture is shown in Figure 6.2 consisting of three parts: analyzing the activity patterns, activity correlation, and the ranking based on the correlation. In the activity analysis, the activity level of Twitter users was analyzed in two different time domains: hour and week. In the hour domain, the average status updates for each hour of the day was calculated. The week domain contains the average status updates for each day of the week.

Activity patterns were analyzed in two time domains: hour and week, and each one was divided into 24 and 7 sub time intervals, respectively. An example of this can be seen in Figure 6.3. On each time interval, the average number of status updates was calculated.

**Figure 6.2.** The architecture consists of three parts: activeness analysis, calculating activeness correlation, and ranking activeness correlation.
(a) The Hour Domain

(b) The Week Domain

**Figure 6.3.** Examples of activity profiles of four users across each time domain.

to represent the activity pattern.

### 6.3.2. Vocabulary Usage Similarity

There are a number of approaches for measuring text similarities: the simple matching coefficient, the Jaccard Coefficient, the Tanimoto Coefficient, Correlation, and Euclidean distance. They all have their own benefits and disadvantages, depending upon which kind of
dataset is considered, but all of the above methods can be applied to find text similarities. To make a valid argument of which one is the most efficient for the task at hand, every method should be compared after being implemented and run on the same dataset. In the study, the Euclidean distance approach was implemented after witnessing its decent performance with a smaller portion of the dataset.

In this part, the content of the status updates was analyzed. The chosen approach to finding social relevance from status updates was divided into three parts, as shown in Figure 6.4. In Part 1, some basic natural language processing techniques were applied in order to

**Figure 6.4.** The status updates similarity model consists of three main parts: preprocessing, weighting vocabulary, and calculating the vocabulary similarity between users and their followers’ vector space model.

...
time it takes to complete by minimizing the feature space. In addition, having stop words
tends to reduce the performance of the system, shown in Figure 6.4.

The next part of the analysis was to weight each vocabulary word: not all words
in the data set are evenly influential, since some words are frequently used in many status
updates and other words are used by only a smaller set of users in their status updates.
TF-IDF is a method used commonly to weight vocabulary:

\[
w_i = \frac{v_{i,j}}{\sum_k v_{k,j}} \cdot \log \left( \frac{|D|}{|\{j : t_i\}|} \right)
\]

TF-IDF gives higher weight to vocabulary that is used frequently in a set of status updates
and gives lower weight to vocabulary that is used frequently throughout the data set.

In the final part, a vector space model (VSM) was created to calculate similarities of
the VSM between users and his/her followers. To create the VSM, first, a bag of vocabulary
words were generated for everyone by combining each user’s and follower’s status updates.
Then a dictionary of words was generated, which included all of the unique vocabulary words
in the data set. Each word was weighted by the equation 8:

\[
W_i = \frac{|F_i|}{|C|}
\]

, where \(|F_i|\) is the frequency of a word \(i\), and \(|C|\) is the total number of status updates. The
result of the vector space model is shown in Equation 9:

\[
V_{1,...,n} = w_1 \cdot W_1, w_2 \cdot W_2, ..., w_n \cdot W_n
\]

After finishing the calculation of the VSM, the vocabulary usage similarity was calculated
between users and their followers by applying the Euclidean distance function. The similarity
score \(SS\):

\[
SS(U_A, U_i) = Euclidean(U_A, U_i)
\]

indicates how many vocabulary words were similar between users and followers.
The final stage of this analysis was to rank the followers according to their similarity score. The next section discusses the results of the two approaches, which were attempting to find the social matrix in order to define social relevance.

6.4. Discussion of Results

Using equation 5 and equation 10, the total activity correlation indexes and vocabulary similarity scores were calculated. The cumulative result of the sample data set is shown in Figure 6.5. On each coordinate, the value went from zero to one: zero indicated the most dissimilarity between users and followers in respective attributes, and a value of one indicated the most similarity. Even though distribution skewed toward one on the y-axis, there

![Figure 6.5](image-url)
was a large accumulation of points on zero. This indicated that there were a large number of followers who used similar vocabulary with their users, but about the same number of followers had very different vocabulary usage from their users, as well.

Analyzing activity correlation indexes on Figure 6.5 revealed that the mean of the distribution was around 0.7: average followers had about a 70% activity correlation with their users.

To evaluate the two attributes that were analyzed in this chapter, the results of the rankings from the activity correlation and vocabulary usage similarity had to be compared against the true rank of social relevance. Social relevance was defined by the number of messages that were exchanged between two different parties: the more messages that were exchanged, the closer the social relevance between the two parties is. This was the base rank to which the results were compared.

The assumption of using the number of interactions as the indicator of social relevance, or base rank, was based on the studies conducted in areas of psychology: socioemotional selective theory[44], and the selective optimization with compensation model[111]. Social interactions change throughout a person’s lifespan from the frequency of interactions with uniform distribution, to a positive skew distribution. The change was explained by the social gain maximization, which is to increase social and emotional gain while reducing emotional risk.

The following steps were taken in the evaluation process, and Figure 6.6 shows an example of a user with 38 followers. The figure shows the calculated similarity score for the total activity correlation index and vocabulary similarity score and fitted lines, as well. The variance of the total activity correlation index was much smaller than the variance of the vocabulary similarity score.

Step 1, the followers were ordered based on the number of messages they received from their user on the $x$-axis. Step 2, scores of activity correlation/vocabulary similarity were plotted. Step 3, linear regression was applied to model the relationship between base ranking and similarity scores, or to show the general trend of the relationship: positive or
Figure 6.6. Applying the linear regression model to fit a line on the similarity score reveals a tendency of the relationship between the calculated similarity score and the status update count.

A positive relationship indicated that if a follower had a high similarity score, he/she tended to receive a high number of status updates. Finally, the slope of the fitted line was calculated to observe the relationship trend.

Table 6.1 shows an overall result of both the activity correlation and the vocabulary similarity

<table>
<thead>
<tr>
<th></th>
<th>Activity Correlation</th>
<th>Vocabulary Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.075042773</td>
<td>0.012846259</td>
</tr>
<tr>
<td>STD</td>
<td>0.180408439</td>
<td>0.450923205</td>
</tr>
<tr>
<td>Max</td>
<td>0.573787607</td>
<td>1.241298145</td>
</tr>
<tr>
<td>Min</td>
<td>-0.377833415</td>
<td>-1.108898947</td>
</tr>
</tbody>
</table>
similarity on the sample data set of size 4,518 and 2,93. Positive mean values in the table indicated that there was a positive correlation between the similarity score and the base rank. The vocabulary similarity’s low mean value and high STD on the table concluded that the activity correlation gave a better measurement than the vocabulary similarity for defining social relevance, even though it was intuitive to assume that people who use similar vocabulary have a high social relevance.

The side by side comparison between activity correlation and vocabulary similarity is shown in Figure (6.7). The distribution of the activity correlation slopes was narrower than the distribution of the vocabulary similarity slopes, which was a good indication that activity correlation is a better approximation of social relevance than vocabulary similarity.

![Distribution of Activity Correlation Slopes](image1.png)  ![Distribution of Vocabulary Similarity Slopes](image2.png)

**Figure 6.7.** Side by side distribution comparison between activity correlation and vocabulary similarity.

6.5. Conclusion

Online social networking has become popular among all generations, being used for both broadcasting and sharing information. Even though it is an efficient method of communication, there are still some privacy issues involving who on the social graph is able to view personal information. Existing social networking sites lack the privacy control needed to manage this beneficially. One approach is to find socially relevant followers with whom users want to share their information. In this preliminary study, non-obvious ways to define
social relevance were examined and it was discovered that activity patterns and vocabulary similarities can be good representatives to define social relevance between users. The result of both attribute studies shows that there is a positive correlation. In future work, more ways to find social relevance will be explored.
CHAPTER 7

DYNAMIC DATA: BEHAVIORAL PATTERNS

In the previous two chapters new ways to capture the strength of a social tie were discussed by using activity correlations and vocabulary correlations. In this chapter, privacy when sharing information on social networks is the main focus. A more granular approach was applied to the problem by identifying behavioral patterns, and model how to select a subset of the social graph to share in an instance. Ideally, users will be able to share their information to a relevant subset within their social graph without making the content public. Determining user behavior patterns was approached in two ways: by identifying shared key words, and by the shared time windows between users. The results showed that the shared time windows method was more promising than using the shared key words method.

The objective of this particular study was to improve the current trend of broadcasting information on OSN sites towards multi-casting information. The difference between the two methods is shown on Figure 7.1: in a broadcast paradigm, when a user makes a status update, all of the user’s connections are notified, as shown in Figure 1(a). On the other hand, only the set of connections who would be interested in the update receives the update in the multi-cast paradigm, shown in Figure 1(b).

To achieve this goal, two approaches were explored: the first approach was shown to not be adaptable enough to work in an OSN environment, while the second approach proved to be much better in defining user behavior. The first approach collected keywords used by both users and the users’ partners and then selected a set of partners who used those common keywords. In the second approach, just the set of partners who had similar behavioral patterns as the users were selected to receive the update. The next section explains the drawbacks of using the first approach.

7.1. Background

After concentrating on identifying features that can represent users’ interests, it was concluded that defining users’ interests based on keywords contained in their status updates
Distribution of information is moving from a broadcast paradigm to a multi-cast paradigm.

is not an efficient way to profile users. The following are some of the reasons that make this approach unattractive:

1. Identifying keywords requires certain knowledge of Natural Language Processing, which is not the most desirable of methods for examining microblogging sites because of its frequent grammatical and spelling errors, brevity, and informal language. One method that did not involve NLP was to treat each word as a tag and do a frequency analysis on these tags, then select the top x tags as keywords that describe the user.
The above-mentioned approach was feasible, however, there were some difficulties. If the approach was applied to users and the users’ friends, each individual often had a different set of keywords to describe him or her, so it generated a curse of dimensionality problem. When a new keyword was added in the feature dimensions, dimensional space increased so fast that the probability of two or more users who used the same keywords decreased rapidly. Therefore, it was very difficult to find common attributes between the users and their friends when this method was used.

The following are some findings from other related research. Golbeck et al., found correlations between vocabulary used in status updates on Twitter and The Big Five Personality Inventory (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism). Subjects in the study were given a test to identify their personalities for these five measures. Features from the subject’s tweets were extracted by the Linguistic Inquiry and Word Count (LIWC), MRC Psycholinguistic Database, and using the general Inquirer database to acquire psychological, relativity, and sentiment features. To predict the personality, Gaussian Processes and ZeroR classification methods were tested, and both generated similar results. The result of the study concluded that the five personality traits were predicted within 11% - 18% of the actual value [70]. Research [115] also showed that personality traits were expressed in tweets as linguistic cues.

Lampe et al., showed that there is a weak correlation between self-descriptive content and number of friends. It was shown that there is a significant difference in the number of friends when comparing whether or not profile fields were being populated. The top five fields affiliated with the largest differences were high school, favorite music, AIM, Birthday, and About Me. [96]

A research conducted by Kosinski et al., [117] concluded that the personality of users can be predicted by only three attributes of counts: followers, following, and list. There were also a lot of research done on prediction problems [46] [98] [110] [64]
7.2. Methodology

To reiterate the objective of the problem, when a user posts status updates the proposed system would send them to only a set of friends to whom the status updates are relevant. This is a better approach than sending the updates to a user’s entire social graph. This way the system protects users’ privacy, and it also reduces the amount of irrelevant status updates that overcast friends’ news feed. Although traditionally this has not been considered to be a serious problem, the popularity of OSN sites as being a medium of communication is growing, which means that it does have to be addressed eventually in the same manner as spam emails are. Thus, the problem remains that when a user posts a status update, how would he or she select a set of friends who are related to the status update or who are interested in reading that status update?

The selected approach to this problem was to find a set of friends from the social graph to whom the user is comfortable to share that specific status update. There were a few things to consider when selecting the set of friends. The privacy control for selecting the set of friends had to be dynamic. This means that when a user decides to update her status, the system had to be able to select a different set of friends depending on the content of the status update. The content of the status update can be about news articles, trending topics on the social network, or routine activities, and it changes time to time depending on the user’s activity for that specific time.

Next everyone’s behavior in a social graph was captured during different time windows. When a user posts a status update during a specific time window, a set of the user’s friends who had the most similar behavior during that specific time window were selected to receive the status update. Therefore, the approach to select a set of friends from a social graph was based on a three part process:

- Identifying behavior during a time window
- Quantifying similarities between user’s and friends’ behavior
- Selecting a set of friends for sending posts within the time window
An assumption here was that people who either worked in the industry, attended grade school and college, or were enjoying their retirement would usually have had a structured schedule and a routine that they followed on a daily and weekly basis.

Behavior is expressed by the content of the status updates. There were specific keywords that were considered common:

1. Total number of post count
2. @ keyword
3. URL keyword
4. # keyword
5. RT keyword
6. NL keyword
7. Average msg length in words
8. SD of the msg length in words

Each feature often indicated a different behavioral feature: the @ symbol (represented a message on Twitter was being tagged with someone’s name), URL symbol (represented a number of links included in text essays), Hashtag (#) symbol (represented a trending topic on Twitter), and RT keywords (represented that one’s status update is being reposted).

Closely analyzing these features revealed behavioral information about the users. For example, frequent usage of the @ symbol indicated that a user liked to have conversations on Twitter by replying to his/her friends’ status updates. Another interpretation could have been that the user commonly tagged his/her friends in status updates that were relevant to the tagged friends in the perspective of the user.

Some interpretations of using URLs in the status updates were that users liked to read news and keep up with current events, so users often posted that information to share it with their followers. However, Twitter usage is not limited to only human users, since some are news curators and spammers which continually attempt to reach new audiences.

---

1The study can be extended to monthly and yearly routines. This particular study only used daily and weekly time windows due to the available data set.
This had the potential to negatively affect the interpretation of URL usage as a behavioral feature.

Hashtags were mainly used for search engines being able to locate and designate keywords towards trending topics. Frequent usage of hashtags could have also indicated a user who actively participated in current events and evolving discussions.

RT was used to re-broadcast someone else’s status update to one’s own social graph. Having frequent RT’s may have represented a connection of common interests or personality between the original poster and the RT user, which further illustrated their behavior.

Aggregating all of the status updates posted on a specific time window was hypothesized to capture individuals’ behavior. Two types of time domains were considered for this study: day and week. The day domain had a length of one day and twenty four hour time windows. On the other hand the week domain had 168 (24\(\text{hoursaday} \times 7\text{daysaweek}\)) time windows. In each time window, eight behavioral features were considered. While the day domain expressed coarse behavior, week domain was for granular behavior.

Figure 7.2 shows seven day time windows from Monday through Sunday, with each day being split into 24 hour windows.

![Figure 7.2. Behavioral signature was calculated on each hour of the day.](image)
After defining the behavior, the behavioral similarity between the user and friends was calculated by the sum of the cosine coefficient of eight features:

\[
\text{sim}(U_t, F_t) = \frac{\sum_{i \in f} U_i \times F_i}{\sqrt{\sum_{i \in f} (U_i)^2} \times \sqrt{\sum_{i \in f} (F_i)^2}}
\]

where \(\text{sim}(U_t, F_t)\) was in range of \([0, 1]\), \(U_t\) and \(F_t\) were feature vectors of a user and a friend on a time window \(t = [0..23]\) and \(f\) was the number of behavioral features. Features @, URL, #, RT, and NL were normalized by the total number of post counts in the time window. The cosine coefficient approach is one of the most commonly used approaches to document text similarity. Advantages of the cosine coefficient approach are that it eliminates absent feature terms and is also capable of handling non-binary values.

The next step was to select a set of friends to whom a given status update should be sent. To do this, an example of the agglomerative clustering, hierarchical clustering approach was applied. The hierarchical clustering algorithm is shown in Algorithm 1. In the context of behavioral similarity, the hierarchical cluster arranged friends of the user into a number of groups. Friends in the same group had similar behavior differences from the user. To proceed with hierarchical clustering, a proximity vector was calculated using Equation 11 in Equation 12.

\[
dis(U_t, F_t) = 1 - \text{sim}(U_t, F_t)
\]

To merge two groups, intergroup similarity was accounted and it was defined by Equation 13.

\[
disG_t(G_A, G_B) = \frac{1}{N_A N_B} \sum_{i \in A} \sum_{j \in B} \text{dis}(i_t, j_t)
\]
It implied that the user was the centroid of the cluster and friends were clustered depending on their behavioral differences from the user’s behavior.

The final stage of the system after building the hierarchical cluster was to find the threshold value that divided the cluster into two sub-clusters. The cluster which had higher behavioral similarity to the user was selected to receive a status update from the user in a given time window. Figure 7.3 shows an example of the final stage of the system in a time window, and it can be seen from the figure that the threshold value of 0.41 separated the cluster into two sub-clusters. The cluster which overall had the most similar behavior to the user was nominated to receive the update in the given time window.

7.3. Experiment Result

To test the hypothesis, the experiment was run on a Twitter data corpus. The data corpus had a total of 4,510,296 million conversation threads (one conversation may have had multiple threads), 477,045 unique conversations, and 296,486 users. A data set of this size was very challenging to study due to limited resources. To study this data, the uniform sampling approach was applied to select a number of sample graphs. The following steps were taken in order to sample a smaller portion of the data corpus:

- Order users by their number friends
- Remove outliers from the users, who are on the top of the list and the bottom of the list using 95% and 45%, respectively.
- Randomly select a user from the rest of the users
- Find the user’s social graph of friends
- Gather all the user’s and his/her friends’ data from the corpus

A sample of ten social graphs with a total of 506 subjects from the corpus were considered for the study. A summary of ten sample data is shown in Table 7.1. Column Grph Size in the table shows the sample size, and columns day and week shows the reduced sample size for day and week domain experiments.
Figure 7.3. Hierarchical clustering was used to split friends into two sets of receiving status updates and not receiving status updates on a time window at thresed value.

The available data presented some challenges in evaluating the performance of the system. For example, sample one (S1) was a social graph of one user who had 32 friends. Each individual in this social graph wrote status updates during different time windows,
such as time of the day and day of the week. It was nearly impossible to find a time window during which all of the individuals in the graph had behavioral data.

Due to this reason, and to accurately measure the performance of the system, friends who did not have behavioral data on a specific time window had to be removed from that experimental run. Column \textit{Rmv} indicates the average number of friends removed from experiment, and column \textit{Exp} shows the average size of the experiment after removing friends who lacked behavioral data.

\begin{table}[h]
\centering
\begin{tabular}{lcccc}
\hline
 & \multicolumn{2}{c}{Day} & \multicolumn{2}{c}{Week} \\
\cline{2-5}
 & Grph Size & Rmv & Exp & Rmv & Exp \\
\hline
S1 & 32.00 & 18.68 & 13.32 & 25.38 & 6.62 \\
S2 & 18.00 & 8.14 & 9.86 & 12.21 & 5.79 \\
S3 & 89.00 & 51.63 & 37.37 & 78.06 & 10.94 \\
S4 & 65.00 & 38.87 & 26.13 & 56.57 & 8.43 \\
S5 & 8.00 & 2.56 & 5.44 & 5.07 & 2.93 \\
S6 & 117.00 & 61.16 & 55.84 & 98.53 & 18.47 \\
S7 & 33.00 & 19.75 & 13.25 & 28.28 & 4.72 \\
S8 & 65.00 & 30.78 & 34.22 & 53.50 & 11.50 \\
S9 & 18.00 & 7.59 & 10.41 & 13.90 & 4.10 \\
S10 & 61.00 & 32.52 & 28.48 & 52.97 & 8.03 \\
\hline
Tot & 506.00 & 271.68 & 234.32 & 424.47 & 81.53 \\
\hline
\end{tabular}
\caption{Basic stat.}
\end{table}

The sample data was further processed into two sets; one was for the day domain study and the other one was for the week domain study. In the day domain study, each user’s data was split into 24 time intervals, and during each interval behavioral features were extracted. On the other hand, there were 168 (7 days × 24 hours) time windows in the week domain study.
As observed in Table 7.1, the amount of data removed from an experiment was higher in the week domain study than in the day domain. It was due to the fact that the probability of two individuals who have behavioral data available on the same time window was \( \left( \frac{1}{168} \right)^2 \), which was much less than the probability of the same event on a day domain of \( \left( \frac{1}{24} \right)^2 \). Therefore, data in week domain was much sparser than the data in the day domain, and less data was available to experiment on.

Given that a user made a status update during a specific time window, behavioral features of both the user and the friends were extracted for the experiment. Then the friends were divided into two groups of similar behavioral features and dissimilar behavioral features.

To evaluate the performance of the method, the result of the experiment was compared with the gold standard. The gold standard was actual information during a given time window of whether the user and any of the friends exchanged any conversations, so the gold standard contained all of the friends who exchanged conversation with the user in each time window.

Using the gold standard, the experiment was evaluated, but there were a few concerns when evaluating the result. First of all, the evaluation did not include the results if a user’s data was missing in a specific time window; this could have happened because of nighttime or work hours. Secondly, friends who had available behavioral data in a specific time window were considered in the result of the evaluation. Finally, there had to be at least three or more friends’s behavioral data in a specific time window.

One way to evaluate the result of the proposed method using behavioral similarity was to apply Pearson’s chi-square test \( (\chi^2) \) to determine whether the result of the proposed method had a significant association with the gold standard. In other words, \( \chi^2 \) was used to test for independence if the significance level was 5%.

\[ H_0: \text{The null hypothesis was that the result of the proposed method and gold standard are independent.} \]

\[ H_a: \text{The alternative hypothesis was that the result of the proposed method and gold standard are not independent or dependent.} \]
Table 7.2 summarized the result of the Pearson’s chi-square test on day and week domains, and the result of the chi-square test was compared with the random set. The comparison result was set up as following: in a given window, the proposed method was applied and it calculated the number of correct and incorrect assignments of positive and negative groups (positive groups were a group of friends who were selected to receive a status update, and the negative group was the opposite). In the day domain, there were 24 such time windows and 168 time windows in the week domain.

The chi-square test was applied on the results from all of the time windows in each domain separately. The result of the chi-square test was summarized in columns $H_0$, P-value and chi-square statistic ($\chi^2$) in the table. The column $H_0$ showed the status of whether the null hypothesis was statistically independent or not. The value 0 indicated to reject the null hypothesis due to the probability ($P-value$) of getting a chi-squared statistic as extreme as ($\chi^2$) is low at the significance level of 5%. In other words, rejecting null hypothesis meant that there was a statistical relationship between the gold standard and the result of the proposed method.

For good measure, the proposed method was compared with a random data set – the result is shown in column Random in Table 7.2. Comparing with the random set demonstrates the worthiness of the proposed method by comparing it with the base case (random set). In the random test, the same number as in the proposed method of random friends were selected in a given window. The chi-squared test was applied on the result of the random test.

It can be observed from Table 7.2 that the result of chi-squared test on day domain clearly indicated that the proposed method performed much better than the random test on selecting a set of friends to send a status update. While seven of ten test cases rejected the null hypothesis, all ten null hypothesis of random set were accepted. The reason that three cases ($S2$, $S5$, and $S9$) of the proposed method failed to reject the null hypotheses was that there was not enough data. Those samples had less than or equal to 10.41 of the experiment size, shown on Table 7.1.
Table 7.2. Chi Square test of independence.

<table>
<thead>
<tr>
<th></th>
<th>Proposed</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Random</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$</td>
<td>P-value</td>
<td>$\chi^2$</td>
<td></td>
<td></td>
<td>$H_0$</td>
<td>P-value</td>
<td>$\chi^2$</td>
<td></td>
</tr>
<tr>
<td>Day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>0.00</td>
<td>0.00</td>
<td>104.48</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.12</td>
<td>76.29</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>1.00</td>
<td>0.33</td>
<td>42.27</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.79</td>
<td>31.68</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>0.00</td>
<td>0.00</td>
<td>148.46</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.91</td>
<td>40.56</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>0.00</td>
<td>0.00</td>
<td>201.50</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.81</td>
<td>55.73</td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>1.00</td>
<td>1.00</td>
<td>22.47</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>25.43</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>0.00</td>
<td>0.00</td>
<td>460.13</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.26</td>
<td>60.11</td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>0.00</td>
<td>0.01</td>
<td>83.66</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.99</td>
<td>34.68</td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>0.00</td>
<td>0.00</td>
<td>203.39</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.94</td>
<td>36.01</td>
<td></td>
</tr>
<tr>
<td>S9</td>
<td>1.00</td>
<td>0.05</td>
<td>64.88</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>25.39</td>
<td></td>
</tr>
<tr>
<td>S10</td>
<td>0.00</td>
<td>0.00</td>
<td>126.85</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.56</td>
<td>57.68</td>
<td></td>
</tr>
<tr>
<td>Week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>1.00</td>
<td>1.00</td>
<td>104.99</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.97</td>
<td>115.36</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>1.00</td>
<td>1.00</td>
<td>59.39</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>47.60</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>1.00</td>
<td>0.11</td>
<td>181.21</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>97.27</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>1.00</td>
<td>0.12</td>
<td>215.53</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>102.80</td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>1.00</td>
<td>1.00</td>
<td>37.41</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>59.72</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>0.00</td>
<td>0.00</td>
<td>465.48</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>127.60</td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>1.00</td>
<td>1.00</td>
<td>76.15</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>71.23</td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>0.00</td>
<td>0.00</td>
<td>251.50</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>81.55</td>
<td></td>
</tr>
<tr>
<td>S9</td>
<td>1.00</td>
<td>1.00</td>
<td>89.91</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>105.92</td>
<td></td>
</tr>
<tr>
<td>S10</td>
<td>1.00</td>
<td>0.78</td>
<td>243.36</td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>174.22</td>
<td></td>
</tr>
</tbody>
</table>

This was also a problem in the week domain tests. The experiment size of the samples were so small that the proposed method did not give good results.

Next, the accuracy, precision, recall, and f-measure of the proposed method with were compared with random sets, and the summary is shown in Table 7.3. On average, the
proposed method performed much better than the base case, and it was even more clear on the day domain. The main problem in this study was that there was not enough data in the week domain. However, when there is enough data the proposed method has a very good potential, which was shown with accuracy of 96 and 8 in the week domain on Table 7.3 with the experiment sizes shown in Table 7.1.

7.4. Discussions

When a user posts a status update, a set of users who have similar behavioral patterns with the user is selected to receive the update. What remains is how to verify that the automatically selected set of partners are actually the ones who should receive the update. To test this, the result of the similarity score was compared against the number of messages which were actually sent from the user to those partners. The result can be verified in two settings: given a day of the week, and given a day and an hour.

In the first setting, there was only the one parameter of a day of the week. In this case, there are 24 hours in the day, and 456 features (19 * 24):

\[
\text{simDay}_d(U, P) = \sum_{h=1}^{24} \sum_{f=1}^{19} \text{sim}_h(U_f, P_f)
\]

In the data set, a number of status updates were sent from the user to each partner on each day. Using this data, the Pearson correlation was applied to see if there was any evidence to support the following hypothesis: if the user has a high behavioral similarity with the partners on a specific day, the user will also be likely to send a post to them. The result of the analysis is shown in Table 7.4. Six of the seven days had a positive correlation.

The day which had a negative correlation was investigated further, and it was found that there were only eleven updates made from the user to the partners. Correlation scores were high on the days with a high number of updates.

After calculating the behavioral similarity of signatures between the user and the partners, the result was visualized on a heat map, shown in Figure 7.4. The figure illustrates similar behavioral patterns between partners and the user. The X-axis in the figure indicates partners, and the Y-axis demonstrates a similarity value for each day of the week. Similarity
<table>
<thead>
<tr>
<th></th>
<th>Proposed</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc   P   R   F</td>
<td>Acc   P   R   F</td>
</tr>
<tr>
<td>Day</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.79  0.33  0.87  0.40</td>
<td>0.55  0.36  0.65  0.41</td>
</tr>
<tr>
<td>S1</td>
<td>0.70  0.38  0.76  0.47</td>
<td>0.65  0.44  0.69  0.52</td>
</tr>
<tr>
<td>S2</td>
<td>0.89  0.10  0.91  0.17</td>
<td>0.50  0.15  0.64  0.24</td>
</tr>
<tr>
<td>S3</td>
<td>0.61  0.22  0.78  0.30</td>
<td>0.51  0.17  0.61  0.26</td>
</tr>
<tr>
<td>S4</td>
<td>0.54  0.71  0.84  0.74</td>
<td>0.67  0.47  0.72  0.54</td>
</tr>
<tr>
<td>S5</td>
<td>0.74  0.10  0.77  0.15</td>
<td>0.50  0.13  0.60  0.21</td>
</tr>
<tr>
<td>S6</td>
<td>0.65  0.25  0.82  0.35</td>
<td>0.57  0.26  0.73  0.37</td>
</tr>
<tr>
<td>S7</td>
<td>0.76  0.13  0.82  0.21</td>
<td>0.52  0.14  0.64  0.23</td>
</tr>
<tr>
<td>S8</td>
<td>0.43  0.73  0.62  0.58</td>
<td>0.54  0.44  0.56  0.48</td>
</tr>
<tr>
<td>S9</td>
<td>0.68  0.27  0.79  0.37</td>
<td>0.53  0.27  0.61  0.35</td>
</tr>
<tr>
<td>Mean</td>
<td>0.68  0.32  0.80  0.37</td>
<td>0.55  0.28  0.65  0.36</td>
</tr>
<tr>
<td>Week</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.66  0.42  0.87  0.52</td>
<td>0.67  0.43  0.85  0.53</td>
</tr>
<tr>
<td>S1</td>
<td>0.61  0.49  0.75  0.51</td>
<td>0.62  0.51  0.73  0.57</td>
</tr>
<tr>
<td>S2</td>
<td>0.82  0.17  0.95  0.27</td>
<td>0.57  0.21  0.85  0.32</td>
</tr>
<tr>
<td>S3</td>
<td>0.67  0.30  0.89  0.39</td>
<td>0.62  0.27  0.81  0.38</td>
</tr>
<tr>
<td>S4</td>
<td>0.54  0.82  0.80  0.76</td>
<td>0.69  0.68  0.82  0.68</td>
</tr>
<tr>
<td>S5</td>
<td>0.81  0.11  0.95  0.18</td>
<td>0.53  0.14  0.86  0.23</td>
</tr>
<tr>
<td>S6</td>
<td>0.61  0.49  0.91  0.54</td>
<td>0.64  0.47  0.93  0.59</td>
</tr>
<tr>
<td>S7</td>
<td>0.76  0.14  0.93  0.23</td>
<td>0.59  0.15  0.81  0.25</td>
</tr>
<tr>
<td>S8</td>
<td>0.52  0.79  0.78  0.73</td>
<td>0.61  0.69  0.83  0.72</td>
</tr>
<tr>
<td>S9</td>
<td>0.65  0.33  0.83  0.40</td>
<td>0.60  0.35  0.86  0.45</td>
</tr>
<tr>
<td>Mean</td>
<td>0.66  0.40  0.87  0.45</td>
<td>0.61  0.39  0.83  0.47</td>
</tr>
</tbody>
</table>

strength is indicated by the intensity of the color, so a darker color means more similarity between the user and that partner. There are some interesting conclusions which can be
Table 7.4. Result of the Pearson correlation analysis shows that six of seven days do support the hypothesis. Using behavioral features of the user in a given day, a set of partners can be selected to receive the status update.

<table>
<thead>
<tr>
<th></th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>-0.032</td>
<td>0.522</td>
<td>0.360</td>
<td>0.212</td>
<td>0.129</td>
<td>0.469</td>
<td>0.573</td>
</tr>
</tbody>
</table>

Table 7.5. Total number of status updates were sent from the user to the partners on each day of the week.

<table>
<thead>
<tr>
<th></th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>11</td>
<td>72</td>
<td>27</td>
<td>33</td>
<td>23</td>
<td>31</td>
<td>54</td>
</tr>
</tbody>
</table>

Figure 7.4. Similarity strength is indicated by the intensity of the color, so a darker color means more similarity between the user and that partner.
drawn if partners’ similarity strength were ordered by the least similar to the most similar on each day of the week, as shown in Figure 7.5. On Tuesday, the user sent status updates to the partners who had similar behavioral patterns because it had the highest density. Another explanation can be that the most amount of updates were sent on Tuesday, shown in Table 7.5. If the same numbers on Friday are viewed, it is significant because there were only 23 updates which were sent from the user, but on the heat map Figure 7.5 Friday had the second highest density of behavioral similarity. Thus, the user was most likely to communicate with the partners who shared similar behavioral patterns on Fridays as well as on Tuesdays.

A similar study for hour windows was conducted - the result is shown on Tables 7.6 and 7.7. 78% of the time, the windows do support the hypothesis.

**Figure 7.5.** The heat map shows a relationship between the user and the partners which have similar behavioral patterns.
Another way to look at the heat map was to order the partners according to their overall similarity with the user, which is shown in Figure 7.6. The last row in the heat map shows the overall similarity of the partners to the user in the order of the least similar to the most similar behavioral pattern. The overall similarity is calculated by summing each day’s similarity score:

\[
sim_{\text{All}}(U_t, P_t) = \sum_{d=1}^{7} \sim_d(U_t, P_t)
\]

The heat map reveals that there were high concentrations of heat on Tuesdays and Fridays, especially closer to the right side of the map. This might have indicated that the partners who have high similarity scores tended to communicate with the users on Tuesdays and Fridays. Heat distribution is the scarcest on Saturdays.

7.4.1. Limitations

Individuals on Twitter follow a specific behavioral pattern of posting status updates, shown in Figure 7.7, so it is observed that in some time windows there are no feature values. This can create an impossible case for this approach to estimate the similarity of the user’s and partners’ features. In Figure 7.7, the subject is most active from midnight to 3; however there are no activities during time intervals of 9 am and 10 am. During these inactive times, a different approach should be applied to estimate similarity.

7.5. Conclusions

In this study, different methods related to finding a socially relevant set of a user’s followers were explored. An assumption was made that social relevance can be defined by the number of messages that are exchanged between two different parties: the more messages that are exchanged, the closer the social relevance between the two parties is. However, the question that remains is what other attributes can help to define social relevance? Two attributes were analyzed here: activity matching and vocabulary usage similarity matching between users and their followers.

The result of the study do support the proposed hypothesis that when a user posts a status update, only a set of the user’s partners can be selected to receive the update. To select
Figure 7.6. The partners of the user are ordered according to their overall behavioral similarity to the user in the order of the least similar to the most similar behavioral pattern. The overall similarity score is shown in the last row in the heat map.

The partners, compare the user’s and the partners’ behavioral features on a specified time window and select the ones who have similar behavioral patterns during that time window. The hypothesis is supported by 85% of the day window and 78% of the hour window.

Further studies need to be done on selecting attributes that indicate behavioral features, and the hypothesis needs to be tested on a larger data set.
Figure 7.7. Distribution of post frequency in a 24 hour window.
Table 7.6. Result of the Pearson correlation analysis shows that 41 of 52 time windows do support the hypothesis. Using behavioral features of the user in a given day and an hour, a set of partners can be selected to receive the status update.

<table>
<thead>
<tr>
<th></th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.025</td>
<td>0.500</td>
<td>-</td>
<td>0.299</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>-0.009</td>
<td>0.657</td>
<td>0.350</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.408</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>0.264</td>
<td>0.355</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>0.105</td>
<td>0.035</td>
<td>-0.038</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.553</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>0.185</td>
<td>0.227</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>0.181</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.052</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.207</td>
<td>-</td>
<td>0.245</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>0.186</td>
<td>-</td>
<td>-</td>
<td>0.332</td>
<td>-0.026</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.060</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>0.198</td>
<td>0.569</td>
<td>-</td>
<td>-</td>
<td>0.138</td>
<td>-</td>
<td>-0.037</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>0.328</td>
<td>0.168</td>
<td>0.135</td>
<td>0.196</td>
<td>-</td>
<td>0.474</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>0.143</td>
<td>-</td>
<td>-</td>
<td>0.157</td>
<td>-</td>
<td>0.194</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>0.273</td>
<td>-</td>
<td>-</td>
<td>0.189</td>
<td>0.181</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>-0.010</td>
<td>-</td>
<td>0.202</td>
<td>0.006</td>
<td>-0.044</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>0.311</td>
<td>-0.040</td>
<td>-</td>
<td>-</td>
<td>-0.042</td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>0.111</td>
<td>-</td>
<td>0.397</td>
<td>-0.029</td>
<td>-</td>
<td>0.748</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.504</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.483</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.705</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>22</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>23</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 7.7. Total number of status updates sent from the user to the partners on each hour of the day.

<table>
<thead>
<tr>
<th></th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>20</td>
<td>-</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>17</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>3</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>14</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>10</td>
<td>-</td>
<td>6</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>17</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>5</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>5</td>
<td>-</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>-</td>
<td>4</td>
<td>1</td>
<td>-</td>
<td>20</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>22</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>23</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
CHAPTER 8

ANALYZING STATIC DATA IN ONLINE SOCIAL NETWORKS

In this chapter, users’ profile information, otherwise known as static data, was collected from the University of North Texas network on Facebook. Interesting behavioral patterns of various demographics were found in the way that users’ revealed personal information on their online profiles. Chapter 9 shows how machine learning approaches can be applied for privacy configuration using these behavioral patterns as shown in this chapter.

8.0.1. Available Data on Facebook

At the time of this research, applications were able to access a database of information using the Facebook API as given in Table 8.1, which could have been considered to be sensitive information for some users. A complete list was available at [62].

8.1. Analysis on Information Revealed

Demographic factors, such as age, education level, and wealth, influence the level of privacy concerns [146]. This section explores how much information is revealed by different demographics.

To find out how much information Facebook users reveal on their profile, a sample of 4,919 Facebook users were analyzed on the University of North Texas network, out of 34,790 registered members. The gender ratio was 35% female and 65% male. The research showed that 75% of the users revealed their education history after high school, shown in Figure 8.1; 70% disclosed their high school’s name; more than 60% posted their favorite movies, music preferences, interests, relationship status, and age; 57% listed books they like; between 45-51% revealed their hometown, their favorite TV shows, activities, and political preference which can be one of libertarian, apathetic, very conservative, conservative, moderate, liberal, or very liberal; and, 21% disclosed the city where they currently resided.

\footnote{http://developers.facebook.com}
Table 8.1. Using the Facebook platform, application developers were able to access users’ personal and social information.

<table>
<thead>
<tr>
<th>Tables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>user</td>
<td>User profile information: first name, last name, birthday, sex, hometown location, current location, political preference, religion, work history, education history, interests, activities, etc.</td>
</tr>
<tr>
<td>friend</td>
<td>All friends of a user. Facebook API method returns list of user IDs (uid).</td>
</tr>
<tr>
<td>group</td>
<td>Groups a user belongs to along with group IDs (gid), names, group types, and descriptions.</td>
</tr>
<tr>
<td>group_member</td>
<td>Member list of a specific group</td>
</tr>
<tr>
<td>event</td>
<td>Upcoming events organized by groups or friends along with that event’s unique ID (eid).</td>
</tr>
<tr>
<td>event_member</td>
<td>Invited members’ status of an event.</td>
</tr>
</tbody>
</table>

In this section, the analysis of data was reported from four categories: gender, age group, relationship status, and political preference. In fact, building a privacy protection system was further investigated in Chapter 9 based on the results shown in Section 9.2. Each category exposed interesting results. Although Facebook did not require that users reveal the information, 62% revealed their age; 64% exposed their relationship status; and 48% showed their political preferences.

In Table 8.2, the data was sorted into subgroups. From 4,919 profiles examined, 65% revealed the user as female and 35% male. 82% of users who were between 20 and 24 years
Figure 8.1. An example of how much personal information was revealed on the UNT social network site, based on 4,919 users.

old were found to reveal their age and 18% of those whose ages fell within one of the groups of 15-19 and 25-29 were likely to share their age. Table 8.2 (Relationship Status) shows that the majority of the users who revealed their relationship were single or in relationships. Table 8.2 (Political Preference) shows that 84% of the users who revealed their political preference were either conservative, moderate, or liberal.

From Figure 2(a), it was found that female users were less likely to reveal their personal and social information when compared to male users. Figure 2(b) shows that users between age 20 and 29 revealed their personal and social information. In addition, in Fig. 3(a) it can be seen that users who are single revealed more information than any other status. Apathetic and libertarian users disclosed more information than any other political preferences as shown in Fig. 3(b).

8.2. Conclusion

It can be seen from the statistical analysis that there were specific behavioral patterns in how different demographics revealed their personal information. The following chapter
Table 8.2. Demographic of data revelation by gender, age, relationship, and politic preference.

<table>
<thead>
<tr>
<th>Category</th>
<th>Group</th>
<th>Revealed %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>65</td>
</tr>
<tr>
<td>Age</td>
<td>15-19</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>20-24</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>25-29</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>30-34</td>
<td>0.5</td>
</tr>
<tr>
<td>Relationship</td>
<td>Single</td>
<td>47</td>
</tr>
<tr>
<td>Status</td>
<td>In Relationship</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Engaged</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Complicated</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Open Relationship</td>
<td>0.4</td>
</tr>
<tr>
<td>Political Preference</td>
<td>Very Liberal</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Liberal</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Conservative</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Very Conservative</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Apathetic</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Libertarian</td>
<td>3</td>
</tr>
</tbody>
</table>

utilizes these patterns in order to apply a default configuration setting for various personal profiles.
(a) Information revealed by different genders.

(b) Information revealed by different age groups.

Figure 8.2. Information revealed based on gender and age.
(a) Information revealed by users in different relationship statuses.

(b) Information revealed by users with different political references.

Figure 8.3. Information revealed based on relationship status and political preference.
CHAPTER 9

DEFAULT PRIVACY CONFIGURATION

9.1. Introduction

Communication among friends, colleagues, and family members has been changing from strictly face-to-face interaction to increasingly include cyberspace, where individuals can receive/send messages, share photos, join groups, read gossip, and meet with strangers. The number of people with Internet access has fostered an environment where real-life social activities transform into online social activities [77]. As with real-life social networks, this online social networking can be clustered into sub-networks with common values. As anyone can join a particular sub-network, this makes OSN sites quite diverse.

Since they do not require face-to-face interaction, OSN sites make it easy to find and meet people. To ensure online interaction takes place, users tend to post personal information such as their actual name, birthday, political preference, religion, relationship status, interests, activities, favorite music, listings of movies, and other information they believe will attract others. This posting provides credibility (through identifying credentials) and suggests areas of compatibility between parties. In addition, the OSN sites allow users to distribute their information through the network efficiently and quickly. For example, users having parties can send event invitations to their networks. Users may learn more about friends on the social network site than they would in face-to-face meetings. Thus, sharing common interests and ideas with friends makes OSN sites an attractive cyber-place to hang out [109]. Since people are more willing to post personal information, it is important to have a mechanism in place which actively protects this personal information unless a user specifically chooses to publicize it.

In this chapter an automatically generated default privacy configuration is proposed to protect the privacy of new Facebook users. The two types of privacy-management systems utilize the statistical modeling based on the analysis from the previous chapter. The initial model used a probabilistic approach, and an improvement was made on the model using
Bayesian Belief Network (BBN) approach. Applying the BBN approach boosted the system accuracy from 75% to 90%.

9.2. Privacy Protection Mechanism

The privacy protection mechanism proposed attempts to mathematically formalize users’ data revelation behavior on social network sites. The preliminary model was developed to facilitate this privacy protection mechanism. Facebook users can configure what information should be available to platform applications. However, Facebook privacy setting is configured after opt-out. In other words, all of a user’s profile information is accessible unless the user knows that specific information must be configured to be inaccessible. To address this issue, a privacy-protection system (PPS) is used to automatically configure a user’s privacy settings based on the user’s profile information. Figure 9.1 shows the model of this privacy-management system. The system uses three components: profile information (PI), privacy manager (PM), and profile zoning (PZ).

Profile Information (PI) contains two types of information: personal and social information. The personal information contains age, relationship status, and political views.
On the other hand, social information consists of hometown, current residency, activities, interests, music, TV shows, movies, books, high school information, education history, profile note, and wall postings. Personal information was selected to be the main feature able to characterize users’ personality and is easily obtained from profile information.

In profile zone (PZ), the PPS divides the user’s profile information into two zones. One zone carries information which is accessible by platform applications. The other zone carries information which can not be accessed by platform applications.

Privacy configuration takes place in the privacy manager (PM). The PM performs its task in two phases. The first phase is sampling the network to discover information — revelation behavior. Sampling should not take place every time a new user joins the network; instead it should happen only after a significant change in network population or a substantial number of users change their personal information. In this study, 4,919 users in the University of North Texas (UNT) network were sampled. In the second phase, the PM configures the user’s privacy setting using revelation matrix and threshold matrix. With this system, users would have to opt-out of privacy rather than opt-in to make all of their information public.

9.2.1. Building Revelation Matrix

The revelation matrix was built by using a statistical analysis on the personal information of the 4,919 UNT users. Personal information was categorized as follows: gender with two subgroups of male and female; age with four subgroups of ages ranging from 15 to 19, 20 to 24, 25 to 29, and 30 and up; relationship status with five subgroups of single (S), in relationship (IR), engaged (E), married (M), and open relationship (OR); and political preference with seven subgroups with very liberal (VL), liberal (Li), moderate (M), conservative (Con), very conservative (VC), apathetic (A), and libertarian (Ln). In each subgroup, the same statistical analysis was applied. Using the male subgroup as an example, the first step was to find all male users from the sample of 4,919 users, and then the percentage of the
users who revealed age, hometown, current residence, etc. were calculated. Equation (16)

\[ R_{i,j} = \frac{n_i}{N_j}. \]

was applied for each subgroup. \( R_{i,j} \) was the percentage of users who were in \( j \) subgroup revealing their feature \( i \), where \( i = \{ \text{Age, Hometown, Cur.resident...}, \text{Wall} \} \), \( j = \{ \text{Male, Female, 15–19,..., Ln} \} \), \( n_i \) and \( N_j \) were the total number of users corresponding to \( i \) and \( j \), respectively.

The data summary after applying Equation (16) is shown on Tables 9.1 and 9.2. Each element of the revelation matrix can also be interpreted as a probability of users revealing profile features based on their personal information.

9.2.2. Building Threshold Matrix

The second step for developing the privacy manager was to build the threshold matrix. The threshold matrix showed what was the average feature revelation for each subgroup. Using the revelation matrix, the threshold was calculated as Equation (17)

\[ T_j = \frac{1}{|F|} \sum_{i \in F} R_{i,j}. \]

, where \( T_j \) was the average profile — the feature revelation for subgroup \( j \) and \( |F| \) was the size of the feature set.

The threshold matrix in Table 9.3 shows that what the average probability was for each subgroup.

9.2.3. Configuring Privacy Settings

To find suitable privacy settings for users’ profile features, the joint probability technique was used on the four main features: age, gender, relationship status, and political preference. Because the statistical analysis was done on the main features with replacement (independent variables), the probability that users reveal their profile features was the product of the main features.

After completing phase 1: building revelation and threshold matrices, the configuration of privacy settings were able to take place. Unlike phase 1, phase 2 took place every time a new user joined a network. The probability of revealing a feature was calculated by the
Table 9.1. Demographic of features revelation by gender and age.

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>Age</td>
<td>0.83</td>
<td>0.51</td>
</tr>
<tr>
<td>Hometown</td>
<td>0.64</td>
<td>0.39</td>
</tr>
<tr>
<td>Cur. resident</td>
<td>0.29</td>
<td>0.16</td>
</tr>
<tr>
<td>Gender</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Relationship</td>
<td>0.85</td>
<td>0.53</td>
</tr>
<tr>
<td>Political view</td>
<td>0.66</td>
<td>0.38</td>
</tr>
<tr>
<td>Activities</td>
<td>0.68</td>
<td>0.40</td>
</tr>
<tr>
<td>Interests</td>
<td>0.84</td>
<td>0.52</td>
</tr>
<tr>
<td>Music</td>
<td>0.88</td>
<td>0.55</td>
</tr>
<tr>
<td>TV shows</td>
<td>0.64</td>
<td>0.43</td>
</tr>
<tr>
<td>Movies</td>
<td>0.85</td>
<td>0.54</td>
</tr>
<tr>
<td>Books</td>
<td>0.75</td>
<td>0.47</td>
</tr>
<tr>
<td>High school</td>
<td>0.92</td>
<td>0.59</td>
</tr>
<tr>
<td>Education</td>
<td>0.98</td>
<td>0.62</td>
</tr>
<tr>
<td>Note</td>
<td>0.21</td>
<td>0.14</td>
</tr>
<tr>
<td>Wall</td>
<td>0.96</td>
<td>0.60</td>
</tr>
</tbody>
</table>

The joint probability of four main features of personal information, and it was compared against the joint probability of the threshold value of given subgroups. If the value was greater than the threshold’s value, the feature was set to be accessible by platform applications Equation (18), otherwise the feature was not accessible Equation (19), where $U_p$ was a set of person $p$’s personal information e.g., $U_1 = \{Male, 24, S, Mo\}$.

\[
\text{if } \left( \prod_{k \in U_p} R_{i,k} \right) > \left( \prod_{k \in U_p} T_k \right), \text{ accessible.}
\]
Table 9.2. Demographic of features revelation by relationship and politic preference.

<table>
<thead>
<tr>
<th></th>
<th>Relationship status</th>
<th>Political preference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>IR</td>
</tr>
<tr>
<td>Age</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Hometown</td>
<td>0.63</td>
<td>0.64</td>
</tr>
<tr>
<td>Cur. resident</td>
<td>0.27</td>
<td>0.28</td>
</tr>
<tr>
<td>Gender</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Relationship</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Political view</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>Activities</td>
<td>0.68</td>
<td>0.67</td>
</tr>
<tr>
<td>Interests</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>Music</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>TV shows</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td>Movies</td>
<td>0.87</td>
<td>0.86</td>
</tr>
<tr>
<td>Books</td>
<td>0.78</td>
<td>0.75</td>
</tr>
<tr>
<td>High school</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>Education</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Note</td>
<td>0.25</td>
<td>0.21</td>
</tr>
<tr>
<td>Wall</td>
<td>0.96</td>
<td>0.96</td>
</tr>
</tbody>
</table>

\[
(19) \quad if \left( \prod_{k \in U_p} R_{i,k} \right) \leq \left( \prod_{k \in U_p} T_k \right) \text{, not accessible.}
\]

The following is an example to show how the automatic configuration of privacy settings works. A new user who is male, 24 years old, single, and has moderate-political preference joins the UNT network. By default, all features are inaccessible. Based on Table 9.1 and 9.2, it is observed that the probability of the user’s revealing age is comparing joint probability (revelation matrix) of 83% (Male), 84% (S), and 84% (Mo) against the joint probability (threshold matrix) of 73% (Male), 73% (20-24), 73% (S), and 77% (Mo). The
Table 9.3. Threshold matrix was built for categories of gender, age, relationship status, and political preference.

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Age</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>15-19</td>
</tr>
<tr>
<td>Threshold</td>
<td>0.73</td>
<td>0.46</td>
<td>20-24</td>
</tr>
<tr>
<td></td>
<td>0.73</td>
<td>0.73</td>
<td>25-29</td>
</tr>
<tr>
<td></td>
<td>0.71</td>
<td>0.73</td>
<td>30-Up</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relationship status</th>
<th>Political preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>IR</td>
</tr>
<tr>
<td>Threshold</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>0.79</td>
</tr>
</tbody>
</table>

result is 59% (revelation matrix) > 30% (threshold matrix), so the user’s age is accessible by the applications.

9.3. Performance

To evaluate the performance of the PPS, it was tested with two sets of data. Based on the dataset, it was found that 3,173 (65%) of the 4,919 users did not provide one or more of the main features (age, relationship status, or political views). One dataset had all 4,919 members without any alteration, and the other one was filtered such that there were only users who had all four main features provided. The performance of the PPS in terms of accuracy rate is shown in Figure 9.2 where the accuracy rate is measured by correctly configured privacy settings of randomly selected users from the dataset. After the system configured the users’ privacy settings for each feature based on their personal information, the configuration was compared with the user’s actual setting. Without filtering the data, the accuracy converged to 70%. On the other hand, with filtered data, the accuracy of the privacy configuration converged to 75%. Even though 65% of the 4,919 users lacked one or more of the personal information parts, the PPS was still able to perform with 70% accuracy.
Figure 9.2. Comparing the accuracy of the privacy management system using a filtered dataset of 1,746 users and an unfiltered dataset of 4,919 users. The system shows notable high error tolerance. 

In other words, with only five percent less than 75%, it indicates that the system had a high error tolerance.

9.4. Conclusion

Social networking boundaries appear to have been pushed in every way possible to allure new users and to keep current users. Open platform gives great flexibility to application developers to be creative and innovative as possible. Even though it gives benefit to both Social Network Sites and to application developers, customers’ private information can be accessed without permission. Not all developers are legitimate. Without careful consideration of privacy management, open platforms result in information harvesting for illegal or unethical purposes. A privacy management system was proposed as a solution to these privacy problems.
The system used a probabilistic approach based on information revelation of users in order to recommend more appropriate privacy settings. The system restricted access to some aspects of users’ personal information unless they consciously chose to make the information available. This experiment showed that this approach was able to achieve 75% accuracy. In addition, its high error tolerance made it a potentially suitable technique for the user content management environment.

9.5. Alternative Approach to Privacy Management

In this section the privacy management model was improved by using the Bayesian Belief Network (BBN). Figure 9.3 shows the architecture of the model, using three components: profile information (PI), privacy manager (PM), and profile zoning (PZ). The model makes the decision (PM) on what profile information/feature (PZ) should be revealed and should not be revealed based on given profile information (PI).

![Figure 9.3](image.png)

**Figure 9.3.** The Privacy Management model was based on three components: PI, PM, and PZ. PM configured user’s profile information into allowed and not allowed to access categories.
Constructing BBN required two steps. Step one involved creating the structure of the BBN using either the TPDA algorithm or the B-Course algorithm. The TPDA algorithm has three phases: drafting, thickening, and thinning. It is proved to be efficient in learning and does not need ordering of the nodes [47, 49]. On the other hand, the B-Course uses a combination of stochastic and greedy search heuristics to explore the very high-dimensional Model spaces [107]. The result of the B-Course algorithm was given in Figure 9.4. The next step was training the BBN, which created the conditional probability tables for each arc in the BBN.

![Complete graph of the learned structure of B-Course.](image)

**Figure 9.4.** Complete graph of the learned structure of B-Course.
9.6. Results of BBN Approach

In order to provide an indication of the accuracy of the proposed approach, a sample of ten queries were conducted to compute the probabilities for a given set of events. The results were compared with the actual values. Two examples of the queries are listed next: 1) If age = 20, gender = male(1), relationship status = in a relationship(2), and political preference = moderate(3), what is the probability of revealing the rest of the features (hometown location, current residency, activities, interest, music, TV shows, movies, books, high school info, education history, note counts, and wall counts)?

The result of the query predicts that, for example, more than 37% of the users will not reveal their current residence; the actual value from the data set presents that 41% indeed did not reveal this feature for the specified group (90% accuracy). The query also indicates that, respectively 95.1% and 96.1% of the users for the query would have revealed their high school information and their education level; for the selected group all the users have revealed such information. The estimates for the remaining features as well as for the other queries present similar results. These results are a clear indication of a correlation between the exposures of some features for a given group of users and indicate the potential of the proposed approach.
10.1. Future of Social Networks

Social networking has become a part of our identity on the Internet, where personal and behavioral information about ourselves can be easily collected from profile information and activities. OSN sites are one place where a vast amount of personal information and behavioral information is gathered through a number of provided services. Technological advances has brought new innovations in the types of services available on these sites; in the era of Web 1.0, social network sites were mainly static pages connected by hyperlinks to other static pages of user profiles. The next progression came with Web 2.0, and OSN sites placed more emphasis on dynamic content such as status updates and photo uploading. Twitter was the pioneer in dynamic content, and soon afterwards the other social networking sites followed suit and integrated dynamic features. The amount of dynamic content generated on social networking sites has been growing exponentially in last few years, and in 2012, 175 million tweets were sent daily on average.

Today there are many more services offered, including an instant messaging service, private messaging, photo uploading, status updating, etc. The purpose of these features are to add value to the overall social network experience in order to encourage users to stay on the site for more time. However, these services do not enable the full potential of OSN sites. In the future there will hopefully be more of an emphasis on interpreting the data collected, using dynamic features such as status updates to play a key role. Status updates can contain a variety of information about the users; some of the common themes of Twitter posts can be

- past, current and future plans
- completed tasks
- interesting online article and video links
- questions and answers
• recommendations and reviews
• links to educational and entertainment sites
• whereabouts
• feelings

As such, tapping into the knowledge of the content can accurately represent the identity of individuals. The progression has to be coincided with advancement of semantic web and natural language processing.

Data generated on a social network is just one aspect of data that can be collected, however. Today there are many more online applications that provide a variety of services such as restaurant ratings, messaging services, document sharing and storing, communication tools, and social networking. Many of these services are free of charge because it is much more valuable to have users data – it can generate profit. There are many free Google services are already available online, such as Google Drive, Google Voice, Google+, Google Reader, Gmail, etc. These are just scraping the surface of many applications that are available. On the same note, Facebook has been diversifying types of services available for Facebook users. On top of social networking, Facebook now offers instant messaging, private messaging, mobile texting and video calling. Gathering data from all of these different resources can provide a lot of granular information about identity, behavior, and social circles of a given user. For example, Personally Identifiable Information (PII) such as gender, age, work and education history, friend circles, invited events, interest of news topics, preferred communication types to contact each friends, etc. can be acquired using these services.

Overall, the types of services available on social networking have diversified to capture a wide range of social data from users. One of the basic services is profile information which provides static content for social networking, and is a requirement for the sign up process. Unfortunately, if users are new to the system, they tend not to modify their privacy configuration [94]. During the period of transition between a novice user and experienced user, personal information on the profile is available to anyone. This dissertation addressed this vulnerability using the privacy knowledge of experienced users to configure the new users’
default privacy settings. A customized default configuration helped to protect accidental information leakage. The system was able to improve from 75% accuracy to 90% accuracy using a BBN approach.

Even though the improvement of the system was able to reach 90% accuracy, more testing needs to be done on a larger data set that has a better representation of a large social network. The study was more aimed toward the feasibility of applying a BBN approach to improve the system. Applying BBN came to into the picture very naturally. The concept of the model was able to capture the probabilistic relationships between revealing a set of profile features and four main features (gender, age, relationship status, and political view). While analyzing the data, there was a distinct tendency of patterns in the way users reveal their profile features, given a set of main features. For instance, if users were male, a teenager, single, and liberal, they were willing to reveal a lot of their profile features; on the other side of the spectrum, conservative and married females in their 30’s were willing to share much less information, especially for the address of current residency. Future work can extend this project to test the system on a large data set that represents a true social network better, and also apply the system on a data set from a different network and compare the results.

Another type of privacy concern in OSN sites is the leakage of information through status updates or a dynamic content generator. The nature of the way a dynamic content generator can be used brings many challenges when trying to provide for privacy because of the fact that the dynamic information can be updated frequently, and the updates can be target to a different set of audiences at each time. To provide privacy for the dynamic information, the privacy management system has to flexible enough to handle a variety of situations.

Approaching this type of a problem requires a clear understanding of the social science of human interactions and psychology of privacy concerns [69, 84, 25], such as how people form social groups, how to define social ties between users, and how the sense of privacy has changed in human history.
Living in a society and maintaining social contact seems to be ordinary tasks done everyday; however, some anthropological studies points out that social living is not actually as simple as it seems to be [55]. A number of studies in neurobiology, biology, and anthropology pointed out that primates developed a large brain (relative to the body mass) just to aid with the complexity of social living [88, 55, 92]. In the study by Jerison, a neurobiologist, he first discovered two distinctive components of the brain; one was responsible for tasks involving body’s physical needs, and the other component was for cognitively complex tasks. Throughout the evolution of the birds and mammals, the second component of their brain has been increasing in size.

Based on this idea, the origin of the social brain hypothesis (SBH) was proposed by Byrne and Whiten in the late 1980’s. The reason for primate’s relatively large brain correlated to complexity of social activities, such as the group size in female members [101], social learning behavior [118], and clique size of grooming [95].

It is natural to simplify the complexity of social ties, though one’s social network is organized into hierarchical social groups from intimate to distant relationships. Personal information is shared among family members and close friends who are in the intimate group while non-sensitive information is shared distant friends. Thus, the types of information shared depends on the trust bond between individuals.

Modeling privacy control for dynamic information on social networks requires a much deeper understanding of human nature and social interactions in a social network. The basis of the approach taken in this dissertation in solving this problem was to quantify social relationships between individuals. The frequency of social interaction, similarity of vocabulary usage, and behavioral patterns of online activities were used in order to estimate a given two people’s social intimacy and social relevance to one another. The study showed promising results to the problem, but further studies will be conducted in the future.
10.2. Limitations of Current Models

In order to channel and filter information to the correct audience, computers must learn to label and categorize generated content. However, currently it is difficult for computer systems to comprehend human subtlety in generated content. This is where Artificial Intelligence (AI) plays a role, as it aims to help computers understand human behavior. However, currently AI is not yet sophisticated enough to understand semantics of human language. Currently the primary function of existing OSN sites is to store user generated content into a database and is retrieved by query. Much more needs to be done in many different fields of AI in order to begin interpreting the content and truly understanding the context of each message.

A subfield of AI is Natural Language Processing (NLP), which delves in computers learning the semantics of language. However it faces many challenges especially in the context of online social networks because users abbreviate their words and do not consistently follow grammatical rules. Additionally, difficulties arise since each word and sentence can have multiple meanings and interpretations. Ultimately, this means that the informality of language in social networking creates a completely new field of study in NLP.

10.2.1. Complexity of Human Social Graphs

Now that a large amount of human interactions are being handled through online social networks, the shortcomings of how computers fail to understand our relationships are evident. Online social networking even created a new area of study in how human interactions take place in society, as a subfield of AI. As computers touch everyday aspects of life, their faults become more apparent, especially in regard to managing privacy. Although human social graphs are very complex, computer models need to be able to understand social relationships. It is difficult for computers to understand social graphs since relationships change with every single human interaction.

Properly functioning models should be able to combine knowledge of human social graphs with interpreted user generated content in order to channel specific information to
correct audiences. Essentially, computers need to think how we do! This entails designing new algorithms and making more advancements in social computing.

In order to reach this goal of being able to correctly deliver information a model needs to be designed which can learn from trial and error, since there is a chance of making errors or delivery to an incorrect audience. Possible approaches to solving this issue is to use reinforcement learning, where a machine designs a model and a user gives feedback frequently to the machine in the beginning. Over time, the machine learns the patterns and a given user does not have to correct it.

10.3. A New Direction for Online Social Networks

To improve the state of online social networks, social network services need to design a User Interface (UI) that captures the greatly varied personalities of its users. When a user posts a message, the service needs to decipher the meaning of the content and be able to recommend an appropriate audience for that specific message. It also needs to be able to give reasoning why the group of people in that audience would enjoy reading the message.

Current standards of filtering by OSN sites such as Facebook hide certain messages without any indication or reasoning behind the logic – this causes discomfort among users as they may feel as though they potentially missed out on information from their friends. No filtering at all in a newsfeed or timeline is overwhelming for users to keep up with their social graph and a large portion of the content is irrelevant to them. The proposed system would show reasoning behind filtering choices and this gives it two advantages. 1) It would allow for reinforcement learning as it could be trained specifically by a user over time, and 2) Users can give direct feedback when it filters things incorrectly and eventually give greater satisfaction to users when using the system.

As an example, if a user posts about a favored team winning a soccer match, the machine model proposed would bring up a few groups of people in the social hierarchy whom are more likely to appreciate reading the message. This would include groups titled with names such as "People who tend to like soccer", "Close Friends (who have not proven to dislike
soccer),” and ”Family.” The user would then be able to modify the groups recommended if necessary without needing to deal with tedious lists of friends.

The proposed system would be able to figure out a specific person’s behavior based on the static and dynamic information given in the past, and use it in both understanding content and delivering it. This would be beneficial for both the user posting a given message and the recipients of that message because this system would give reasoning behind every filtering choice it makes. It also would promote social bonds between users because it would show how these people are related to each other through shared content, and even allow users to make new discoveries about people in their social graph. Some users may initially find this to be invasive (or overly ”smart”) but over time it would prove to be an invaluable asset in the fluidity of social relationships and be far more trustworthy overall.
CHAPTER 11

CONCLUSION

Being able to broadcast to a large audience without any barriers via online social networks has proven to be a very powerful tool for society. This idea of an open flow of information is idealized by Mark Zuckerberg, founder of Facebook [138]. The idea itself challenges the old way of delivering news, where information being broadcasted is tightly controlled by a few media giants. Social media is the main reason that many public demonstrations were able to take place against regimes in Egypt, Tunisia, Algiers, Bahrain, and Iran. It actually made differences in these countries. In one sense it can eradicate barriers of delivering news without discrimination and help to increase the openness and transparency of government [39].

Online social networks are a revolution in human communication. It has become a part of our every day life though we had never had any technology readily available to communicate with our social groups in such a broad manner. A vast amount of information from nearly everyone is also readily available to be discovered. However, existing social network sites lack the understanding of human nature in a social context. Online social networking, as we know now, has not naturally evolved to match the ways in which we have evolved to communicate in real life with our network of friends. It lacks the understanding of social science in the way we build relationships and share information with our network of friends.

Socializing and sharing are parts of human nature that we have adopted throughout the evolutionary time period to survive by building alliances to be protected; on the other hand socializing may also mean to seek out information from others for our own benefit. A sense of privacy was developed as an evolutionary by-product of social intelligence [69]. In any context of society, privacy has to be a part of the system that is there to protect oneself from others.
Online social networking sites are places where personal information is readily available to others, which creates vulnerabilities that can be exploited by others. The mechanism to protect ourselves from others in real life was to create a social network of hierarchical social groups. When there is information to be shared, a decision is made to decide with whom the information should be shared. The decision is based on the constraints that sharing this information would not jeopardize the perception of oneself to others.

The wide extent that information can reach has actually become a negative aspect of online social networking when the content of the information is about personal matters. In this situation, people choose the safer route of not sharing that information at all with their social network. If they are unaware of this safety concern, they unintentionally broadcast that information publicly and become victims of unexpected consequences.

Considering the fact that almost one in seven people of world population uses an online social networking site has made it a frontier of a new era in communication, and it has been a "Wild West" ride. Online social networking has come to a point that distinctions between the word friend and the word stranger have blurred into one identity and the differentiation between personal and public information is distorted.

In recent years, there have been some positive steps taken toward protecting privacy. The prime examples of this are privacy settings and the creation of social groups. The predictions in this dissertation on the future of social network is that there will be an organic integration of privacy controls when a user shares information, the content of information will be emphasized more on dynamic data, and that the available data will be processed to make sense of social ties between friends.

The contributions of this dissertation towards the future of social networks are to understand the social structure of communicating, to design social matrices to capture social ties, and to reduce the burden of configuring privacy settings.
APPENDIX A

USER GUIDE FOR DEVELOPING A TWITTER APPLICATION
This user guide is intended for beginners who are researchers or application developers. I focus on researchers who want to collect dataset for their research from Twitter network. In addition to the researchers, anyone who wants to begin developing application using Twitter API can use this guide. In this document, I assume the developers\(^1\) have basic knowledge of PHP. If you have any scripting language experience, you can learn PHP without much difficulty. Because this guide is intended for beginners, I try to cover a wide range of topics, such as introduction to Twitter and setting up a Linux server.

A.1. What is Twitter?

Twitter is founded by Jack Dorsey, Biz Stone and Evan Williams of podcasting company Odeo as research and development project in March 2006, and it was initially used internally by Odeo’s employees. In October 2006, twitter became a product of Obvious and rapidly gained popularity. It was awarded the 2007 South by Southwest Web Award in the blog category. In April 2007, Obvious formed the service as a separate entity under the name Twitter, Inc.,. Twitter is used commonly in Japan and has a high number of users, despite the user interface being completely in English. As a result, Twitter created a version of Twitter for Japanese users in April 2008. Japanese is now the second most used language on Twitter. Unlike the US service, the Japanese service is advertising supported.

On Twitter, users can write a message within 140 characters answering the question "what are you doing?". There is no limit on the number of messages that can be posted. Unlike traditional online social networking (OSN) sites, users change their status every day. Some of the users change it more frequently than others. Information can be extracted from users’ statuses that can be valuable to find the users’ behavior.

If we consider Twitter network as a graph, it is a directional graph, which consists of notes and directional edge between nodes Figure A.1. I use term “Twitter-users” to describe

---

\(^1\)I refer researchers and application developers as developers for clarification.
Figure A.1. An example of a user’s social graph on Twitter. The user is marked with dark gray color, followers are blue, and followees are orange. Edges between two identities are their relationship directions.

A.1.1. Twitter Data

Twitter is very open network to collect data. Majority of the users are willing to reveal their status updates, so people can keep up with their daily activities. It can give great opportunity for researchers in many fields to understand users’ online behavior. In this section,
I show what kind of information can be extracted using Twitter API [137]. There are three main categories of data: social graph, user profile, and status. Data in user’s social graph contains the user’s followers and followees IDs. User profile contains information shown in Table B.5. Most of the profile features are self explanatory, but some may need clarification. Feature ”description” is Twitter-user’s biography. Feature ”protected” has either true or false value. If it is true, user’s profile and status updates are protected. Feature ”utc_offset” is required because every timestamp on Twitter network is converted in Greenwich Mean Time (GMT). Next category of data is status Figure A.2. Some of the status updates contain data in feature ”in_reply_to_status_id”, which means that these status updates are written to respond to someone else’s status update. Feature ”in_reply_to_status_id” contains repliee’s status id.

It is important to understand to what kind of data you can get and how it can be used. You need to keep in mind that what you want to do with this data. Storing data you never use will take space on your disk.

A.2. Setting up the Server

If you have experience with setting up and maintaining any Linux box, this should be easy for you. Otherwise, I cover the how to setup your LAMP server. LAMP stands for Linux, Apache, MySQL, and PHP/Perl. Hardware requirement for running a LAMP server, you need to have a system with at least 256MB RAM. If you have less than this, you will likely to encounter problems caused by insufficient memory. Because you need MySQL to store your data, it is good to have enough memory.

There are plenty of Linux distributions you can select from. If you want know more about your Linux distributions, you can find them at [53]. For this guide, we concentrate on setting up Debian GNU/Linux distribution, which is one of the commonly used ones. If you do not have much experience with Linux distribution, It is better to use popular distribution because you can find more recourse about that distribution on the Internet. You can follow the following steps to set up your server.

\footnote{Repliee implies that someone whom a user is replying to.}
Table A.1. All Twitter users have their profiles. This table shows profile features that can be accessed by the Twitter API.

<table>
<thead>
<tr>
<th>Profile Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>User’s assigned unique ID</td>
</tr>
<tr>
<td>name</td>
<td>User name</td>
</tr>
<tr>
<td>screen_name</td>
<td>User name for twitter account</td>
</tr>
<tr>
<td>location</td>
<td>User’s location</td>
</tr>
<tr>
<td>description</td>
<td>User’s description</td>
</tr>
<tr>
<td>profile_image_url</td>
<td>Link to user’s background image</td>
</tr>
<tr>
<td>url</td>
<td>Link to user’s website</td>
</tr>
<tr>
<td>protected</td>
<td>Privacy setting</td>
</tr>
<tr>
<td>followers_count</td>
<td>Count for number of followers</td>
</tr>
<tr>
<td>profile_background_color</td>
<td>User’s preference for background color</td>
</tr>
<tr>
<td>profile_text_color</td>
<td>User’s preference for text color</td>
</tr>
<tr>
<td>profile_link_color</td>
<td>User’s preference for link color</td>
</tr>
<tr>
<td>profile_sidebar_fill_color</td>
<td>User’s preference for sidebar color</td>
</tr>
<tr>
<td>profile_sidebar_border_color</td>
<td>User’s preference for sidebar border color</td>
</tr>
<tr>
<td>friends_count</td>
<td>Count for number of friends/ followee</td>
</tr>
<tr>
<td>created_at</td>
<td>Account created date</td>
</tr>
<tr>
<td>favourites_count</td>
<td>Count for user’s marked favorite–status updates</td>
</tr>
<tr>
<td>utc_offset</td>
<td>Time difference between user’s timezone and GMT in seconds</td>
</tr>
<tr>
<td>time_zone</td>
<td>User’s timezone</td>
</tr>
<tr>
<td>profile_background_image_url</td>
<td>Custom background image</td>
</tr>
<tr>
<td>profile_background_tile</td>
<td>Custom background tile setting</td>
</tr>
<tr>
<td>statuses_count</td>
<td>Count for number of status updates posted</td>
</tr>
<tr>
<td>notifications</td>
<td>Enables device notifications for updates from the specified user</td>
</tr>
</tbody>
</table>
Table A.2. When users update their status, the following status features are also been created. If a status is written to respond to someone else’s status update, in_reply_to_status features are filled with appropriate information.

<table>
<thead>
<tr>
<th>Status Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>created_at</td>
<td>Status created timestamp</td>
</tr>
<tr>
<td>id</td>
<td>Unique ID for status update</td>
</tr>
<tr>
<td>text</td>
<td>Content of the status update</td>
</tr>
<tr>
<td>source</td>
<td>Interface used to make the update</td>
</tr>
<tr>
<td>truncated</td>
<td></td>
</tr>
<tr>
<td>in_reply_to_status_id</td>
<td>Repliee’s status id</td>
</tr>
<tr>
<td>in_reply_to_user_id</td>
<td>Repliee’s user id</td>
</tr>
<tr>
<td>favorited</td>
<td>Status marked as favorite or not</td>
</tr>
<tr>
<td>in_reply_to_screen_name</td>
<td>Repliee’s screen name</td>
</tr>
</tbody>
</table>

(1) **Installing your Debian GNU/Linux distribution**

Debian GNU/Linux is distributed freely over the Internet. If you have the Internet access during installation, you can do netinst. Otherwise, you need to have complete image of the Debian on CD/DVD.

(2) **Updating your Debian GNU/Linux**

Before proceeding to install, update the necessary packages with this command.

```bash
>> apt-get install update
```

(3) **Installing Apache and PHP**

Apache is one of the widely used web servers for Linux based servers. Latest stable versions of PHP is version 5. The following command installs both apache and php 5.

```bash
>> apt-get install apache2 php5 libapache2-mod-php5
```
You can find your apache configuration file at: `/etc/apache2/apache2.conf` and web folder at `/var/www`. To check your installation, you can create a `test.php` file in your web folder and insert `phpinfo()` PHP function.

```bash
>> nano /var/www/test.php
```

The `nano` command open up an editor insert the following text and save.

```php
# test.php
<?php phpinfo(); ?>
```

Now, go to `http://your.ip.address/test.php` link on your browser, you should see your php configuration and default settings.

(4) **Installing MySQL Database Server**

You can store your data in a database or files. Advantage of using database is that it takes care of competing recourse problem if you have multiple programs accessing same data. On the other hand, you need to handle competing recourse problem if decide to use files to store your data. My recommendation is using database to store data. The following command installs MySQL servier and its client

```bash
>> apt-get install mysql-server mysql-client php5-mysql
```

The configuration file of mysql is located at: `/etc/mysql/my.cnf`

**Creating users to use MySQL and Changing Root Password**

By default mysql creates user as root and runs with no passport. You might need to change the root password. To change Root Password

```bash
>> mysql -u root
mysql> USE mysql;
mysql> UPDATE user SET Password=PASSWORD('new-password') WHERE user='root';
mysql> FLUSH PRIVILEGES;
```
If you want to access your database from PHP script, you should create a user. For your security, you should never use your root password to connect to MySQL database. Next step, you learn how to maintain your database using control panel, such as webmin or phpMyAdmin.

(5) **PhpMyAdmin Installation**  
PhpMyAdmin is web based database management and administration software and easy to install and configure under apache.

```
>> apt-get install phpmyadmin
```

The phpmyadmin configuration file is located at: `/etc/phpmyadmin` folder. To set up under Apache, you need to include the following line in `/etc/apache2/apache2.conf`.

```
Include /etc/phpmyadmin/apache.conf
```

You need to restart your apache server to make sure your last configuration takes place by

```
>> /etc/init.d/apache2 restart
```

You can access to your phpMyAdmin by opening `http://domain/phpmyadmin` page on your browser.

A.3. Storing Data

There are two way you can store your data on your system. You can store it in files or in database. Storing data in file may be appropriate if data is small and your system has insufficient memory. Otherwise, storing data in database can simplifies a lot of work in terms of postprocessing because there are many query commands that can be used to search in your data. If you want to use files to store your data, you may need to implement a lot programs to do search, which is already available in MySQL. You can go to [108] for further information.

When you collecting data from third party servers, you will receive number of unexpected result and error messages. In order to handle this, you should have preprocessing technique to make sure your received data has your expected format. Otherwise, it is very
difficult to find this kind of errors in your data. In addition to preprocessing technique, you also need to have notification system that notify you when there is unexpected error. The notification system can work in two ways when there is an error. It terminates your program and report you about the error by dumping the result in log file. Errors can be handled in the program using exceptions. If a specific error occurs often and you know what to do when it occurs, you can catch the error and handle it accordingly [140].

A.4. Important Information

Before you start developing your applications use Twitter API, you should know basic information about the API.

A.4.1. What’s an API?

Using an API (Application Programming Interface), a application can retrieve or modify data on third party servers. Twitter API provides variety of methods for features you can see on Twitter website. Twitter API is an interface layer for making applications, websites, widgets, and other projects that interact with Twitter. Applications talk to the Twitter API over HTTP.

A.4.2. The API is entirely HTTP-based

Twitter API require a GET request when you use methods that retrieve data from the Twitter. Twitter API require POST for submitting, changing, or destroying data. You can also use a DELETE request to destroy data. API Methods that require a particular HTTP method will return an error if you do not make your request with the correct method. HTTP Response Codes are useful to find cause of an error. HTTP Response Codes are explained in Section FAQ.

A.4.3. The API is a RESTful resource

The Twitter API is modeled after design principles of representational state transfer (REST). It is your preference to request data in different formats because the API supports XML, JSON, RSS, and Atom syndication formats. You select your format by simply changing a request extension in the format. Not all methods support all these formats.
A.4.4. **Parameters have certain expectations**

Some API methods take optional or requisite parameters. When your making requests with parameters, you need to keep in mind the following:

1. Parameter values should be converted to UTF-8 and URL encoded.
2. The page parameter begins at 1, not 0.

There are two special parameters in the Twitter API:

- **callback**: Used only when requesting JSON formatted responses, this parameter wraps your response in a callback method of your choice. For example, appending &callback=myFancyFunction to your request will result in a response body of: myFancyFunction(...). Callbacks may only contain alphanumeric characters and underscores; any invalid characters will be stripped.
- **suppress_response_codes**: If this parameter is present, all responses will be returned with a 200 OK status code - even errors. This parameter exists to accommodate Flash and JavaScript applications running in browsers that intercept all non-200 responses. If used, it’s then the job of the client to determine error states by parsing the response body. Use with caution, as those error messages may change.

Some API methods will return different results based on HTTP headers sent by the client. Using both a parameter and an HTTP header, you can control behavior.

A.4.5. **There are pagination limits**

There is a limit on number of request you can make to Twitter. If your application requires more requests then limit, you may request for whitelisting, such that your application is allowed 20,000 requests per hour.

**Rest API Limit**: Client applications may request up to 3,200 statuses via the page and count parameters for timeline REST API methods. If you reach request limit, you will get result in a reply with a status code of 200 and an empty result in the format requested.

**Search API Limit**: Client applications may request up to 1,500 statuses via the page and rpp parameters for the search method. The response to a request exceeding this limit will
be a status code of 200 and an empty result in the format requested. Twitter also restricts the size of the search index by placing a date limit on the updates it allows you to search.

A.4.6. *cURL and Twitter API*

If your system has curl (and it should!), you can play around the Twitter API [131]. Here are some examples you can try on the command line:

- Get the public timeline in RSS format, unauthenticated:
  
  ```
  >>> curl http://twitter.com/statuses/public\_timeline.rss
  ```

- Get updates from users you follow in XML, authenticated:
  
  ```
  >>> curl -u username:password
  ```

- Get the friends timeline in XML format, unauthenticated:
  
  ```
  >>> http://twitter.com/statuses/friends\_timeline.xml
  ```

- See just the headers for that last request:
  
  ```
  >>> curl --head -u username:password http://twitter.com/statuses/
  friends\_timeline.xml
  ```

- Post a status update and get the resulting status back as JSON:
  
  ```
  >>> curl -u username:password -d status="your message here"
  http://twitter.com/statuses/update.json
  ```

It is a good way to test what the result of Twitter methods look like.

A.4.7. *Twitter API*

The Twitter API consists of two parts: the Twitter REST API methods and the Search API methods. Using the Twitter REST API, you can update timelines, status data, and user information. The Search API methods provide interaction with Twitter Search data trend. You can find detailed information on these methods at [137]
A.5. Crawler Architecture

After successfully setting up the system and understanding the Twitter API, it is time to design the crawler architecture, shown in Figure A.2. The architecture consists of four parts: manager, recovery, crawler, and data storage. The main part of the architecture is manager, which hourly checks on a crawler to decide whether to initiate recovery or crawler part of the system.

![Crawler Architecture Diagram](image)

**Figure A.2.** Crawler architecture consists of four parts: manager, recovery, crawler, and data storage.

A good system must have recovery procedure to be able to handle unexpected situations that causes the crawler to crush. It is especially important when dealing with systems that have unpredictable behaviors, such as having timeout, reaching request limitations, returning unexpected data, etc. In the case of crawler crush, the recovery part is initiated. The first step is to read a log file, and if the cause of the crush can be resolved, initiate the programs that can fix the problem. Otherwise, the recovery document the crush for programmers to later look at. The recovery part is to handle these situations and keep the system functioning smoothly.

The crawler part handles collection data. There are five components in this part: profile collection (Profile), status update collection (Status), network sampling (Sampling...
Method), queue (User IDs), communication to Twitter Server (Twitter API). Profile and status update collections are handled by two separate programs. From a user’s profile, user IDs of the user’s friends can be collected, and depending on a sampling method next user IDs are queued. There are a wide range of network sampling methods can be employed to select the next user profile to collect. In Section B, I propose a network sampling using Kolmogorov complexity. The last part of the architecture is to store the data, received rom Twitter in the database.

A.6. FAQ

Q:: Is Twitter API stable?
A:: Twitter API is relatively new, so changes has been made frequently. This can cause a lot of problems when you developing an application on top of the Twitter API, so it is important to keep yourself up to date about changes [136].

Q:: What request method should I use for my application?
A:: Recently, they have made change in method APIs can request. They used to allow both GET and POST methods. Right now, they only allow GET method. If you try to use POST method, it will be returned with ”bad request”.

Q:: What does blacklisted means?
A:: A developer’s account and/or IP address blocked by Twitter when the developer made more than allotment of requests. Contacting with Twitter may resolve this issue if the is some misunderstandings.

Q:: How do I keep from running into the rate limit??
A:: Two ways you can solve this problem: caching and rate limiting by active user. Caching API responses on you server helps your application run faster and to reduce making redundant requests. Rate limiting by active user.

Q:: What does this HTTP status code means?
A::

- 200 OK: Success!
- 304 Not Modified: There was no new data to return.
• 400 Bad Request: The request was invalid. An accompanying error message will explain why. This is the status code will be returned during rate limiting.
• 401 Not Authorized: Authentication credentials were missing or incorrect.
• 403 Forbidden: The request is understood, but it has been refused. An accompanying error message will explain why.
• 404 Not Found: The URI requested is invalid or the resource requested, such as a user, does not exists.
• 406 Not Acceptable: Returned by the Search API when an invalid format is specified in the request.
• 500 Internal Server Error: Something is broken. Please post to the group so the Twitter team can investigate.
• 502 Bad Gateway: Twitter is down or being upgraded.
• 503 Service Unavailable: The Twitter servers are up, but overloaded with requests. Try again later. The search and trend methods use this to indicate when you are being rate limited.

A.7. Conclusion

Twitter is a online social networking site. In this network, users answer to one question ”What are you doing?” as many and as frequent as they want to, so the followers can keep up with their users’ statuses. Data from this network can reveal interesting information about users behavior. The data is accessible for developers and researchers by the Twitter API. In this user guide, I cover topics that can be most beneficial to researchers who wants to collect data from Twitter for their study.
APPENDIX B

APPLYING KOLMOGOROV COMPLEXITY IN NETWORK SAMPLING
One of the initial problems can be encountered when doing research on online social networking (OSN) is how to collect an unbiased data set. This is a problem most of the time overlooked due to resource limitations, and many researches are preferably conducted on already available data. The main reason is that OSN sites tend be very large, so it is implausible to collect the whole network. A feasible approach to this problem is to sample the network for a representative network, which has relatively similar network properties. There are a number of algorithms for sampling network problems. Each one has its own benefits and costs.

B.1. Related Work

**Node sampling:** In this sampling method, nodes are selected uniformly at random from the whole network. The representative network consists of the nodes sampled and their edges. *Disadvantage* of this model is that it does not accurately capture power-law degree distributions. Node sampling can also lead to biased estimates for clustering and path length properties. To randomly select nodes, it is required to have original network prior to start sampling. It is implausible for large networks.

**Edge sampling:** It is also known as Link sampling. In this sampling method, edges are selected uniformly at random from the original network. Each selected edges and their two incident nodes are added to sample network. *Disadvantage* is bias toward high degree nodes or hubs. Large degree nodes have high probability of getting selected in the sample network. However, this property can also lead to high accuracy in path length distribution. Another disadvantage is loss of structural properties: clustering and connectivity, so the result can be a disconnected and sparse sample network. Like node sampling, it requires to have original network prior to start sampling.
**Topology-based sampling:** The following are some examples of topology-based sampling:

1. Snowball sampling: Snowball sampling algorithm is based on breath first search algorithm with randomly selected seed node. It is not appropriate technique for using limited resource to collect large networks.

2. Random-walk sampling: In simple random-walk sampling, each neighboring nodes has uniform probably of being selected as a next node. Alternative random-walk sampling uses some probability distribution for selecting the next node. *Advantage* is that sampling network is based on original network’s connectivity.

3. Forest fire sampling: Hybrid of snowball sampling and random-walk sampling is forest fire sampling algorithm. A seed node is randomly selected, and the fraction of outgoing links with their nodes from the seed node is burned. The fraction can depend on the application. *Advantage* in practice, this technique can generate good representative of the original network [Sampling from Large Graphs]. It does not require knowledge of the entire network.

**B.2. Algorithm Requirement**

To adopt one of these sampling methods, the following requirements needs to meet:

**No knowledge of original network:** A sampling algorithm does not have prior knowledge of the whole network. Many sampling methods make assumption that they have complete network (all nodes and edges). However, in the reality it is implausible to have the complete network because size of OSN is massive. The question is how to select a node from neighboring nodes.

**Preserve topology of the original network: local clustering coefficient:** The sampling algorithm needs to keep the topology the original network. It is important feature for social graph analysis.
**Use cumulative knowledge:** The sampling algorithm needs to have cumulative knowledge of sampled nodes and makes decision to select its next node based on this knowledge.

**Sample:** Sample size that can represent the original network.

Sampling representative and unbiased network is an important issue to be addressed because many machine learning models are depends on the data set. Data collection process for sampling task can use crawlers, and it is a recommended option for networks as large as Facebook and Twitter. However, it is implausible to collect all data from Facebook or Twitter; Facebook has over 1 billion user accounts and there 600 status updates are posted in every second on Twitter. Depending on sampling algorithm and initial node used for the crawler, underlying true network property may differ from sampled network. Sampling based on uniform/non-uniform random node selection, random node-edge sampling, snowball sampling, degree of user activity based sampling, and specific attribute based sampling are some of the competing techniques for sampling large networks. Each one of these methods has their advantages and disadvantages; I need to do further investigation on methods that preserve topology of underlying social graph. It is one of the three data types I am using in my model.

**B.3. Comparing Network Properties**

Accuracy of sampling algorithms can be defined by how well sampled network preserve some key properties of the original network. To test competing algorithms, network properties of sampled network and the original network needs to be compared to find the best one. The following are some of the properties that can be used to merit sampling algorithms. (Measure goodness of sampling based on the following categories)

- Average degree (in-degree and/or out-degree)
- Average path length
- Density: \(2 \times |E|/(|V| \times (|V| - 1))\)
• Clustering coefficient: measurement of a specific nodes neighboring cluster nodes degree.
• Diameter
• Size of weakly connected component (wcc): if there is an undirected edge from u to v or from v to u.
• Size of strongly connected component (scc): if there is an directed edge from u to v and v to u.
• Number of reachable nodes within hops.
• Densification: number of edges verses number nodes

B.4. Kolmogorov Complexity Approach for Sampling

In this project, I apply kolmogorov complexity for sampling a large online social network. It is also known as descriptive analysis. The main concept is that finding the shortest description of a given input. For example, description of the given input ”aaaabbbbccccddd” can be 4a*4b*4c*4d. In the description there are 4 terms, or the complexity of the input is 4. This is just one of many possible description of the input. Depending on an algorithm complexity is different.

Another application of Kolmogorov complexity can be finding randomness in an input. Let’s assume the input is a sequence of binary numbers. Applying a Kolmogorov complexity algorithm on the sequence of binary numbers can expresses randomness of the input sequence. Complexity or randomness of binary number consisting of all ones or zeros should be smaller than the complexity of binary number consisting of random ones and zeros sequence.

Kolmogorov complexity can express randomness in a given binary sequence. This concept can be applied in sampling online social network/graph. Figure B.1 represents an example of online social graph $G = V, E$, where $V$ is set of vertices/nodes and $E$ is set of edges. Edges of vertices are not visible in the graph. In other words, $V$ is set of online social network users and $E$ connection between the users. This is a preliminary study, so we do not consider connection between nodes. All the sub graphs in Figure B.1 are imaginary clusters.
Figure B.1. Example of a social network graph. In the graph there are six sub graphs marked by purple, and each one has a source node marked by blue node. There can be multiple sub-subgraphs $A_i$ with in sub graph $B$ with property:

$$B_1 \cap B_2 \cap \ldots \cap B_n = \emptyset,$$

and each sub graph can contain sub-sub graphs. All sub-sub graphs can have same number of nodes ($|A_1| = |A_1| = \ldots = |A_n|$). The sub-sub graphs also have same property as sub graph:

$$A_1 \cap A_2 \cap \ldots \cap A_n = \emptyset,$$

Selecting source nodes is another research topic that needs to be in-depth studied, so it is no further explored in this paper. For simplicity, we make assumption that all source nodes are pre-determined with the following identities: no two-source nodes are located in a same sub-sub graph. Each source node belongs to only one sub-sub graphs. All the nodes are in a sub-sub graph are adjutant nodes of a source node.
B.4.1. Definition

The Kolmogorov complexity of a string $x$ is the length of the smallest program that outputs $x$, relative to some model of computation. That is,

\begin{equation}
C_f(x) = \min_p |p| : f(p) = x, \text{some } f
\end{equation}

Informally, Kolmogorov complexity is about finding notion of randomness.

B.5. Data Set

Data set consists of 8,801 users’ profile. It is collected from Twitter in Spring 2000. It is one of the most popular online social networking sites with registered 105 million users, so it can be a good sample for social networking study. Initiation of harvesting 8,801 users’ profiles originates from a single source node, so it may not good representation of the Twitter social network.

Connectivity graph of 8,801 users can be considered not enough connectivity information, such as small degree and high density. In order to work around this, we generate syntactic connectivity for the users. All the users are randomly assigned into $n$ number of sub graphs, and within each sub graph users are further divided into $m$ sub-sub graphs. Size of each sub graph is randomly selected within range of 5-50%. Size of each sub-sub graph is 10% of its sub graph’s size, so all sub-sub graphs within a same sub graph have same size.

In the data set, each user has profile information, which is shown in the following table with its default value and description.

B.6. Methodology

Kolmogorov complexity can express randomness of given input. High complexity indicates there is high randomness in the input. In this project, we used Lempel-Ziv compression algorithm for calculating kolmogorov complexity.

First step is converting users’ profile attributes into binary form. We only interested in whether the profile attributes are changed from their default values or not. If the attributes are changed, binary value of one (1) is assigned. Otherwise, zero (0) is assigned.
Table B.1. All Twitter users have their profiles. This table shows profile features that can be accessed by the Twitter API.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Default Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>uid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>location</td>
<td>Empty</td>
<td>The attribute is filled by the user</td>
</tr>
<tr>
<td>description</td>
<td>Empty</td>
<td>The attribute is filled by the user</td>
</tr>
<tr>
<td>profile_image_url</td>
<td>default_profile_normal.png</td>
<td>Profile image of the user</td>
</tr>
<tr>
<td>url</td>
<td>Empty</td>
<td>Outside of link to the user’s website</td>
</tr>
<tr>
<td>protected</td>
<td>FALSE</td>
<td>By default privacy setting</td>
</tr>
<tr>
<td>bg_color</td>
<td>9ae4e8</td>
<td>Background color</td>
</tr>
<tr>
<td>text_color</td>
<td>0</td>
<td>Text color</td>
</tr>
<tr>
<td>link_color</td>
<td>0000FF</td>
<td>Link color</td>
</tr>
<tr>
<td>sidebar_bg_color</td>
<td>e0ff92</td>
<td>Sidebar background color</td>
</tr>
<tr>
<td>sidebar_border_color</td>
<td>87bc44</td>
<td>Sidebar border color</td>
</tr>
<tr>
<td>following_count</td>
<td>0</td>
<td>Number of users, the user following</td>
</tr>
<tr>
<td>followers_count</td>
<td>0</td>
<td>Number of users following the user</td>
</tr>
<tr>
<td>favourites_count</td>
<td>0</td>
<td>Favorite post count</td>
</tr>
<tr>
<td>utc_offset</td>
<td>3600</td>
<td>Time zone in seconds</td>
</tr>
<tr>
<td>bg_image</td>
<td>Empty</td>
<td>Background image</td>
</tr>
<tr>
<td>bg_tile</td>
<td>FALSE</td>
<td>Background title</td>
</tr>
<tr>
<td>statuses_count</td>
<td>0</td>
<td>Number of status the user posts</td>
</tr>
</tbody>
</table>

Second step is splitting users into $n$ sub graphs. The number of split ($n$) is randomly selected within $[10, 30]$. Size of each sub graph is randomly generated within range of 5-50% from average size of sub graph. $N$ sub graphs are further divided into $m$ sub-sub graph with size of $|n| \times 10\%$. 
Next step is calculating each attribute’s Kolmogorov complexity of first n sub-sub graphs. Mean complexity of the n sub-sub graphs is used as a threshold. Selecting initial source nodes are vital for good representative sample of the whole graph.

Using threshold, the sampling algorithm makes a decision whether to sample further within the sub graph or not. To make the decision we use majority vote with each attribute has one vote. If more than 50% of the attributes vote to stop, the sampling algorithm ceases further sampling nodes connected reachable from a given source node and jumps to another source node (selecting the next source is not in the domain of this study). Voting is based on comparing Kolmogorov complexity of next sub-sub graphs with threshold complexity. The sampling algorithm stops sampling within the sub graph if more than 50% of the next sub-sub graphs attributes’ complexity is less than threshold.

B.7. Experiment

The result of an experimental is shown on the table. Total population of the graph is 8,480. The total population is divided into 21 sub graphs (SG), and each sub graph is split into 10 sub-sub graphs (SSG). Each sub-sub graphs is 10% of its sub graph. Four different cases of vote percentages policy is applied: 50%, 60%, 70%, 80%, respectively. From the last row of the experimental study we can observe that as more restricted rule is applied, the size of the sample increases. In the implementation, policy can be adjusted. If the sampling algorithm is being applied on small social network, restrict voting policy should be applied. However, on a large graph, less restrictive policy might be appropriate.

B.8. Conclusion

All academic researchers use data set for testing models and for studying population. However, we are not always certain that the data set is a good representative of the total population. It is not feasible to collect a large data, such as online social network: Facebook and Twitter. There are 500 million and 105 million users on Facebook and Twitter, respectively. These are two major social networking sites on the Internet today. To model and study such a large network, we need to sample the network with representative data set.
Table B.2. Vote count (VC) shows number of attributes vote to stop sampling further in the sub graph. Using the sampling algorithm sub graphs are reduced to smaller sample size (N').

<table>
<thead>
<tr>
<th>SG</th>
<th>SIZE</th>
<th>50% SSG VC N'</th>
<th>60% SSG VC N'</th>
<th>70% SSG VC N'</th>
<th>80% SSG VC N'</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>280</td>
<td>9 4 280</td>
<td>9 4 280</td>
<td>9 4 280</td>
<td>9 4 280</td>
</tr>
<tr>
<td>2</td>
<td>390</td>
<td>2 10 78</td>
<td>3 12 117</td>
<td>3 12 117</td>
<td>9 8 390</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>2 10 100</td>
<td>3 13 150</td>
<td>3 13 150</td>
<td>6 16 300</td>
</tr>
<tr>
<td>4</td>
<td>410</td>
<td>2 11 82</td>
<td>2 11 82</td>
<td>3 13 123</td>
<td>9 11 410</td>
</tr>
<tr>
<td>5</td>
<td>330</td>
<td>5 10 165</td>
<td>9 10 330</td>
<td>9 10 330</td>
<td>9 10 330</td>
</tr>
<tr>
<td>6</td>
<td>350</td>
<td>2 9 70</td>
<td>4 13 140</td>
<td>4 13 140</td>
<td>9 13 350</td>
</tr>
<tr>
<td>7</td>
<td>500</td>
<td>2 13 100</td>
<td>2 13 100</td>
<td>2 13 100</td>
<td>9 9 500</td>
</tr>
<tr>
<td>8</td>
<td>490</td>
<td>2 9 98</td>
<td>3 11 147</td>
<td>4 14 196</td>
<td>4 14 196</td>
</tr>
<tr>
<td>9</td>
<td>510</td>
<td>2 12 102</td>
<td>2 12 102</td>
<td>2 12 102</td>
<td>3 15 153</td>
</tr>
<tr>
<td>10</td>
<td>450</td>
<td>2 12 90</td>
<td>2 12 90</td>
<td>2 12 90</td>
<td>9 8 450</td>
</tr>
<tr>
<td>11</td>
<td>450</td>
<td>3 10 135</td>
<td>5 13 225</td>
<td>5 13 225</td>
<td>9 11 450</td>
</tr>
<tr>
<td>12</td>
<td>500</td>
<td>2 14 100</td>
<td>2 14 100</td>
<td>2 14 100</td>
<td>2 14 100</td>
</tr>
<tr>
<td>13</td>
<td>570</td>
<td>2 12 114</td>
<td>2 12 114</td>
<td>2 12 114</td>
<td>8 15 456</td>
</tr>
<tr>
<td>14</td>
<td>390</td>
<td>2 14 78</td>
<td>2 14 78</td>
<td>2 14 78</td>
<td>2 14 78</td>
</tr>
<tr>
<td>15</td>
<td>330</td>
<td>2 10 66</td>
<td>9 8 330</td>
<td>9 8 330</td>
<td>9 8 330</td>
</tr>
<tr>
<td>16</td>
<td>260</td>
<td>9 1 260</td>
<td>9 1 260</td>
<td>9 1 260</td>
<td>9 1 260</td>
</tr>
<tr>
<td>17</td>
<td>390</td>
<td>2 13 78</td>
<td>2 13 78</td>
<td>2 13 78</td>
<td>9 10 390</td>
</tr>
<tr>
<td>18</td>
<td>370</td>
<td>3 11 111</td>
<td>3 11 111</td>
<td>4 12 148</td>
<td>9 8 370</td>
</tr>
<tr>
<td>19</td>
<td>250</td>
<td>9 6 250</td>
<td>9 6 250</td>
<td>9 6 250</td>
<td>9 6 250</td>
</tr>
<tr>
<td>20</td>
<td>350</td>
<td>3 9 105</td>
<td>4 12 140</td>
<td>4 12 140</td>
<td>9 8 350</td>
</tr>
<tr>
<td>21</td>
<td>410</td>
<td>2 11 82</td>
<td>2 11 82</td>
<td>5 13 205</td>
<td>6 15 246</td>
</tr>
</tbody>
</table>

| 8480 | 2544 | 3306 | 3556 | 6639 |
In this paper we build a sampling model based on combination Kolmogorov complexity and majority voting algorithm. The sampling algorithm is tested on a data set, combination of real twitter profiles and syntactically generated graph topology. Using 50% voting policy, the sampling algorithm was able to reduce the original data set by 70%. Whole data set is reduced by 60%, 70%, and 80% for voting policies of 60%, 70%, and 80%, respectively. Voting policy should be adjusted for size of whole population. In a large network, less restricted voting policy is appropriate, such that source nodes are selected from wide range of sub graphs. However, restrictive voting policy is appropriate for sampling from a small network.

We made following assumptions to simplify the problem: nodes found in a same sub graph tend to be homogeneous. We also made an assumption that initiating source nodes for the algorithm are given prior to start sampling. In the even of jumping to a new source node, the new source node is provided in the sampling algorithm.

Further experimental studies needs to be done on a large network with known topology. It needs to be verified whether sampled network preserve its original topology, or not. We can test sampled network with the original based on the following graph properties: average degree (in-degree and/or out-degree), average path length, density \((2 \times |E|)/(|V| \times (|V| - 1))\), clustering coefficient (measurement of a specific node’s neighboring cluster nodes degree), diameter, size of weakly connected component (wcc), size of strongly connected component (scc), number of reachable nodes within hops, densification (number of edges verses number nodes).
APPENDIX C

GROUP RECOMMENDATION SYSTEM FOR FACEBOOK
Online social networking has become a part of our everyday lives, and one of the popular online social network (SN) sites on the Internet is Facebook, where users communicate with their friends, join to groups, create groups, play games, and make friends around the world. Also, the vast number of groups are created for different causes and beliefs. However, overwhelming number of groups in one category causes difficulties for users to select a right group to join. To solve this problem, group recommendation system (GRS) is introduced, and it use combination of hierarchical clustering technique and decision tree. Facebook SN groups can be profiled based on their members’ profile. The experiment result showed that GRS was able to make 73% accurate recommendation.

C.1. Introduction

Face-to-face, voice, email, and video communications are traditional medium of interaction between friends, family, and relatives. The traditional medium takes place when two parties had already shared some form of common value: interest, region, family bond, trust, or knowledge of each other. Although, on online social network (SN) two parties initiate communication without the common values between them, they still can freely share their personal information with each other [22]. In the virtual world, joining or creating groups and making friends are a click of a button, which makes online social networking sites, such as Friendster, MySpace, Hi5, and Facebook more and more popular and diverse each day [143]. Therefore, online SN’s advantages are user friendliness and flexible in cyberspace where users can communicate with others and create and join groups as their wishes. Even though flexibility of online SN brings diversity in cyberspace, it can also lead to uncertainty.

The social network data from the University of North Texas (UNT) SN was studied as a sample for the research. There are 10 main group types, such as business, common interest, entertainment & arts, geography, music, etc. Six of them have over 500 groups, and four of them have range between 61 and 354 groups in each. It is overwhelming to find a
group that fits a user’s personality. The study concentrates on identifying inherent groups’ characteristics on SN, so that group recommendation system (GRS) was develop to help the user to select the most suitable group to join.

Groups were created to support and discuss causes, beliefs, fun activities, sports, science, and technology. In addition, some of the groups have absolutely no meaningful purpose, but just for fun. The research shows that the groups are self-organized, such that users with similar characteristics, which distinguishes one group from others. The members’ profile features, such as time zone, age, gender, religion, political view, etc, so members of the group contributes to their group identity. In other words, the group members’ characteristics shape characteristic of the group.

Main Contribution: Group Recommendation System (GRS) can be used to classify social network groups (SNGs). Even though groups consist of members with different characteristics and behaviors, which can be defined by their profile features, as their group size grow, they tend to attract people with similar characteristics [139]. To make accurate group recommendation, hierarchical clustering is used to remove members whose characteristics are not quite relevant with majority in the group. After removing noise in each group, decision tree is built as the engine of our GRS. In this paper, how decision tree can be applied not only to classifying SNGs, but also used to find value of features that distinguish one group from another. GRS can be a solution to online SN problem with the overwhelming number of groups are created on SN sites because anyone can create groups. Having too many groups in one particular type can bring concern on how to find a group that has members with whom one shares common values. If more and more members share common values, the group will grow in size and have better relationship. A good GRS can be a solution to problem of a large number of abandoned groups on social network. Thus, GRS can be a vital feature for social network sites.

The rest of the chapter is organized as follows. Section 2 discusses related work done on social network. Section 3 describe the architecture and framework of GRS. Section 4
presents the performance of GRS. The chapter is concluded with a summary and an outlook on the future work

C.2. Related Work

There has been an extensive number of research efforts focused around modeling individual and group behaviors and structure, but here only a sample set of those research works is referenced. Many researches on social networking have been done in mathematics, physics, information science, and computer science based on properties, such as small-world, network transitivity or clustering, degree distributions, and density [65, 66, 68, 85, 99]. From research in statistics, Hoff et al. developed class models to find probability of a relationship of between parties, if positions of the parties are known on a network [83]. Backstrom et al. has done very interesting research on finding growth of network and tendency of an individual joining a group depends on structure of a group [35].

C.3. Methodology

This section covers data collection process, noise removal using hierarchical clustering, and data analysis to construct decision tree. Figure C.1 shows basic architecture of the group recommendation system (GRS). It consists of three components: i) profile feature extraction, ii) classification engine, and iii) final recommendation.

C.3.1. Facebook API

The dataset we used in this research was collected using Facebook Platform. Facebook launched its API to public in May 2007 to attract web application developers. The API is available in multiple programming languages: PHP, Java, Perl, Python, Ruby on, C, C++, etc. Since Facebook and Microsoft became partners, Microsoft has launched developer tools in its Visual Studio Express and Propfly. The Facebook Platform is REST-based interface that gives developers access to vast amount of users’ profile information.

Using this interface, we had access to student accounts in which privacy setting was configured to allow access to its network (default setting). In our research we used University of North Texas (UNT) social network on Facebook. During this research we were able to
access 1580 users’ accounts. From the accounts, we collected users’ profile information, friend connections, and groups where they belong to. For our analysis, we selected 17 groups from common interest groups on UNT SN. Table C.1 shows detailed information of the groups.

C.3.2. Profile Features

The first step of group recommendation system is to analyze and to identify the features which capture the trend of a user in terms of its interest, social connection, basic information such as age, sex, wall count, notes count and many such features.

Fifteen features were extracted to characterize a group member on Facebook: Time Zone - location of the member, Age, Gender, Relationship Status, Political View, Activities, Interest, Music, TV shows, Movies, Books, Affiliations - number of networks a member
Table C.1. Information of 17 common interest groups on UNT social network including their subtype categories, number of members, and description.

<table>
<thead>
<tr>
<th>Group</th>
<th>Subtype</th>
<th>Group Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>Friends</td>
<td>12</td>
<td>Friends group for one is going abroad</td>
</tr>
<tr>
<td>G2</td>
<td>Politic</td>
<td>169</td>
<td>Campaign for running student body</td>
</tr>
<tr>
<td>G3</td>
<td>Languages</td>
<td>10</td>
<td>Spanish learners</td>
</tr>
<tr>
<td>G4</td>
<td>Beliefs &amp; causes</td>
<td>46</td>
<td>Campaign for homecoming king and queen</td>
</tr>
<tr>
<td>G5</td>
<td>Beauty</td>
<td>12</td>
<td>Wearing same pants everyday</td>
</tr>
<tr>
<td>G6</td>
<td>Beliefs &amp; causes</td>
<td>41</td>
<td>Friends group</td>
</tr>
<tr>
<td>G7</td>
<td>Food &amp; Drink</td>
<td>57</td>
<td>Lovers of Asian food restaurant</td>
</tr>
<tr>
<td>G8</td>
<td>Religion &amp; Spirituality</td>
<td>42</td>
<td>Learning about God</td>
</tr>
<tr>
<td>G9</td>
<td>Age</td>
<td>22</td>
<td>Friends group</td>
</tr>
<tr>
<td>G10</td>
<td>Activities</td>
<td>40</td>
<td>People who play clarinets</td>
</tr>
<tr>
<td>G11</td>
<td>Sexuality</td>
<td>319</td>
<td>Against gay marriage</td>
</tr>
<tr>
<td>G12</td>
<td>Beliefs &amp; causes</td>
<td>86</td>
<td>Friends group</td>
</tr>
<tr>
<td>G13</td>
<td>Sexuality</td>
<td>36</td>
<td>People who thinks fishnet is fetish</td>
</tr>
<tr>
<td>G14</td>
<td>Activities</td>
<td>179</td>
<td>People who dislike early morning classes</td>
</tr>
<tr>
<td>G15</td>
<td>Politics</td>
<td>195</td>
<td>Group for democrats</td>
</tr>
<tr>
<td>G16</td>
<td>Hobbies &amp; Crafts</td>
<td>33</td>
<td>People who enjoys Half-Life (PC game)</td>
</tr>
<tr>
<td>G17</td>
<td>Politics</td>
<td>281</td>
<td>Not a Bush fan</td>
</tr>
</tbody>
</table>

belongs to, Note counts - number of member’s note for any visitors, Wall counts - visitor’s note for member’s page, Number of Friends - number of friends in the group.

The result of the analysis of 17 groups is shown in Figure C.2 illustrates gender ratio, age distribution, and political view in 17 groups. It is also useful to draw parallel attention between Table 1 and Figure C.2. G1 is a friend group, and majority of the members are Female, age between 20 and 24, and 33% do not share their political preference. Same 33% are moderate. These properties identify G1. Same way we can interpret all 17 groups.
Female members are majority in G1 (friends group), G4 (campaign for homecoming king and queen), G7 (Asian food lovers), G10 (clarinet players), G13 (people who likes fishnet), and G17 (Not Bush fan). At same time, majority of G17 consider themselves as liberal. Figure 2(b) shows that majority of all groups are members between age 20 and 24. Figure 2(c) and 2(d) illustrates that majority of G3 (spanish learners), G5 (wearing same pants everyday), G7 (Asian food lovers), G8 (religions group), G10 (clarinet players), G12 (friends group), G16 (PC gamers) did reveal their political preference.

As we can see that using this property, we can construct a decision tree to make better group selection for Facebook users. From Figure 2(a),

C.3.3. Similarity Inference

One of the frequently used techniques to find similarity between nodes in multidimensional space is hierarchical clustering analysis. To infer similarity between members, we use Euclidian distance [119]. Clustering takes place in the following steps for each group:

1. normalizing data (each feature value = [0, 1]
2. computing a distance matrix to calculate similarities among all pairs of members based on Equation 23:

\[
d_{r,s} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_{r,i} - x_{s,i})^2},
\]

where \(d\) is the similarity between nodes \(r\) and \(s\), \(N\) is number of dimensions or number of profile-features, and \(x\) is value at a given dimension.

3. using unweighted pair-group method using arithmetic averages (UPGMA) on distance matrix to generate hierarchical cluster tree as given by Equation 24:

\[
d(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} \text{dist}(x_{r,i}, x_{s,j}),
\]

where \(n_r\) is number of cluster in \(r\), \(n_s\) is number of cluster in \(s\), \(x_{r,i}\) is the \(i\)th object in cluster \(r\), and \(x_{s,i}\) is the \(i\)th object in cluster \(s\). Equation 24 finds average distance between all pairs in given two clusters \(s\) and \(r\).
Next step is to calculate clustering coefficient to find the cutoff point such that noise can be reduced.
C.3.4. Clustering Coefficient

Each group has a unique characteristic, which differentiates it from others, yet some members within the same group may have different profiles. As these differences grow to some extent, these members emerge as an inevitable "noise" for clustering.

To detect and mitigate this noise thus the group is strongly characterized by core members who establish innermost part of the group, we introduce the clustering coefficient (C), which is given by Equation 25:

\[
C = \frac{N_{R_i}}{R_i},
\]

where \( R_i \) is the normalized Euclidean distance from the center of member \( i \), given by Equation 26 hence \( R_i = [0, 1] \), and \( N_k \) is the normalized number of members within distance \( k \) from the center, given by Equation 27 and hence \( N_k = [0, 1] \).

\[
R_i = \frac{r_i}{\arg\max_j (r_j)},
\]

where \( r_i \) is the distance from the center of member \( i \) and \( i = 1, 2, 3, \ldots, M \).

\[
N_k = \frac{n_k}{M},
\]

where \( n_k \) is the number of members within distance \( k \) from the center, and \( M \) is the total number of members in the group.

To reduce the noise in the group, we retain only members whose distances from the center are less than and equal to \( R_x \) as shown in Figure C.3, where \( R_x \) is the distance at which clustering coefficient reaches the maximum.

C.3.5. Decision Tree

The nature of group recommendation system (GRS) is classification type problem. Based on a user’s profile features, GRS finds the most suitable groups for a user. One solution to classification type problem is decision tree algorithm, based on binary recursive partitioning. There are number of splitting rules: Gini, Twoing, and Deviance [43]. To find better result we integrated each of splitting rule to GRS. However, test showed no significant
Figure C.3. An example of $C$ vs $R_i$ plot for finding $R^x$. This plot illustrates the cutoff distance $R^x$ which is the corresponding distance from the center at which the clustering coefficient is maximum and gradually decreasing to 1 at $R_i = 1$. As the clustering coefficient starts to decrease, the sparseness of the outer circular members increases more rapidly since the denominator starts to dominate (greater than numerator) according to Equation 25. Hence the outer circular members or members who have distance from the center greater than $R^x$ to be considered as noise and removed.

Improvement in accuracy, which means that final tree does not depend on what splitting rule is used to construct the tree [43]. The main goal of these splitting algorithms is to find the best split of data with maximum homogeneity on each side. Each recursive iteration purifies data until the algorithm reaches to terminal nodes (classes).

Binary tree consists of parent node $t_p$ and child nodes of $t_l$ and $t_r$. To define maximum homogeneity of child node, we introduce impurity function $i(t)$, so maximum homogeneity of $t_l$ and $t_r$ nodes is equal to the maximum change in impurity function $\Delta i(t)$ (given by Equation 28), which shows that splitting rule go through all variable values to find the best split question $x_i < x_j^R$, so that maximum $\Delta i(t)$ is found:

145
\[
\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r),
\]

where \(P_l\) and \(P_r\) are probabilities of left and right nodes, respectively. Thus, maximum impurity is solved on each recursion step and given by Equation 29:

\[
\max_{x_j \leq x_j, j = 1 \ldots M} [i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r)],
\]

where \(x_j\) is variable \(j\), \(x_j^R\) is the best possible variable \(x_j\) to split, \(M\) is number of variables.

C.4. Result

In this research we developed group recommendation system (GRS) using hierarchical construct and decision trees. To evaluate the performance of GRS, we used 50% of data for training and other 50% for testing. For testing, we selected labeled members and clustered those using GRS. Accuracy rate is measured by the ratio of correct clustered members to total testing members. Figure 4 compares accuracy of GRS with clustering and without clustering for noise removal. Average accuracy without clustering was 64%. Meanwhile, after removing noise from each group using clustering coefficient method, average accuracy improved to 73%. In other words, average accuracy improved by 9%. In addition, 32% of 1580 members or 343 members were found to be noise and eliminated.

![Figure C.4](image)

**Figure C.4.** Accuracy comparison of GRS with and without clustering where the accuracy is improved by 9% with clustering.
C.5. Conclusion and Future Work

It is challenging to find a suitable group to join on SN, especially networks as big as MySpace and Facebook. Until now, online social networking has no sign of slowing down. While Facebook has 42 million users as of October 2007, there are 67 million active users as of February 2008. It has been doubling its size in every six months. To improve quality of service for Facebook users, we developed GRS to find the most suitable group to join by matching users’ profiles with groups’ identity. The system is built using combination of hierarchical clustering and decision tree. After removing noise, we achieved 9% average accuracy improvement over without removing noise and average accuracy of 73%. Nature of decision tree is well suited for generating list of most favorable groups for user. In our future work, we will improve the GRS by listing a certain number of most suitable groups according to the users’ profile. Tree figure on Figure C.1 illustrates that once the most suited group is found, other nodes in same sub–tree or neighbor share similarity with the most suited group. This property can be vital to find list of suitable groups. The main concept behind the GRS can be used in many different applications. One is information distribution system based on profile features of users. As social networking community expands exponentially, it will become a challenge to distribute right information to a right person. We need to have a methodology to shape flooding information to user from his/her friends, groups, and network. If we know identity of the user’s groups, we can ensure the user to receive information he/she prefers. Another research area can be explored is targeted-advertising [48] to individuals on social network site. Many advertising technique are already implemented, such as Amazone based on users’ search keywords and Google Adsense based on context around its advertising banner. In addition, Markov random field technique has emerged as useful tool to value network customer [54].
BIBLIOGRAPHY


[22] Lada A. Adamic, Orkut Buyukkokten, and Eytan Adar, A social network caught in the web, First Monday 8 (2003). 137

[23] Anne Adams, Multimedia information changes the whole privacy ballgame, Proceedings of the tenth conference on Computers, freedom and privacy: challenging the assumptions (New York, NY, USA), CFP ’00, ACM, 2000, pp. 25–32. 16


[27] Denise Anthony, Tristan Henderson, and David Kotz, Privacy in location aware computing environments, IEEE Pervasive 6 (2007), no. 4, 64–72. 32


[58] ______, *The social brain hypothesis*, brain 9 (1998), 10. 15


[67] Eric Gilbert and Karrie Karahalios, *Predicting tie strength with social media*, CHI (Boston, Massachusetts, USA), 2009. 36


[78] Jianming He, Wesley Chu, and Zhenyu Liu, *Inferring privacy information from social networks*, Intelligence and Security Informatics (Sharad Mehrotra, Daniel D. Zeng, Hsinchun Chen, Bhavani Thuraisingham, and Fei-Yue Wang, eds.), Lecture Notes in


[90] Eija Kaasinen, *User needs for location-aware mobile services*, Personal Ubiquitous Comput. 7 (2003), no. 1, 70–79. 36


[110] Salma Noor and Kirk Martinez, *Using social data as context for making recommendations: an ontology based approach*, Proceedings of the 1st Workshop on Context,


