Master Thesis
Computer Science
Thesis no: MCS-2010-21
June, 2010

Predicting friendship levels in Online Social Networks

Waqar Ahmad
Asim Riaz

School of Computing
Blekinge Institute of Technology
Box 520
SE – 372 25 Ronneby
Sweden

This thesis is submitted to the School of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:

Authors:
Waqar Ahmad
E-mail: waqarzahoor@gmail.com
Asim Riaz
E-mail: asim.riaz@yahoo.com

University advisors:
Dr. Henric Johnson, PhD
Department of Systems and Software Engineering
Dr. Niklas Lavesson, PhD
Department of Systems and Software Engineering

School of Computing
Blekinge Institute of Technology
Box 520
SE – 372 25 Ronneby
Sweden

Internet: www.bth.se/com
Phone: +46 457 38 50 00
Fax: + 46 457 271 25

ABSTRACT

Context: Online social networks such as Facebook, Twitter, and MySpace have become the preferred interaction, entertainment and socializing facility on the Internet. However, these social network services also bring privacy issues in more limelight than ever. Several privacy leakage problems are highlighted in the literature with a variety of suggested countermeasures. Most of these measures further add complexity and management overhead for the user. One ignored aspect with the architecture of online social networks is that they do not offer any mechanism to calculate the strength of relationship between individuals. This information is quite useful to identify possible privacy threats.

Objectives: In this study, we identify users' privacy concerns and their satisfaction regarding privacy control measures provided by online social networks. Furthermore, this study explores data mining techniques to predict the levels/intensity of friendship in online social networks. This study also proposes a technique to utilize predicted friendship levels for privacy preservation in a semi-automatic privacy framework.

Methods: An online survey is conducted to analyze Facebook users' concerns as well as their interaction behavior with their good friends. On the basis of survey results, an experiment is performed to justify practical demonstration of data mining phases.

Results: We found that users are concerned to save their private data. As a precautionary measure, they restrain to show their private information on Facebook due to privacy leakage fears. Additionally, individuals also perform some actions which they also feel as privacy vulnerability. This study further identifies that the importance of interaction type varies while communication. This research also discovered, "mutual friends" and "profile visits", the two non-interaction based estimation metrics. Finally, this study also found an excellent performance of J48 and Naïve Bayes algorithms to classify friendship levels.

Conclusions: The users are not satisfied with the privacy measures provided by the online social networks. We establish that the online social networks should offer a privacy mechanism which does not require a lot of privacy control effort from the users. This study also concludes that factors such as current status, interaction type need to be considered with the interaction count method in order to improve its performance. Furthermore, data mining classification algorithms are tailor-made for the prediction of friendship levels.

Keywords: Online Social Network, Friendship Levels, Privacy Concerns, Data Mining

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The online social networks (OSNs) represent an emerging area which also brings many challenges and research opportunities besides numerous socializing facilities for individuals. OSNs try to imitate real life social networks on the Internet and hence support interaction and communication among people. The purpose of this thesis is twofold; highlighting privacy related issues confronted in OSNs with their latest solutions and classifying friendship levels in an OSN by using data mining techniques. This chapter serves as an introduction to the research challenges as well as the strategies which are utilized to achieve these research goals. The remainder of the chapter is organized as follows: Section 1.1 introduces the problem domain. Section 1.2 and Section 1.3 defines research goals and questions, respectively. In Section 1.4, the methodology of research is discussed in detail and, since the research process may raise some validity threats, a discussion about the identified potential threats is carried out in Section 1.5. A review of the related work is provided in Section 1.6 whereas Section 1.7 summarizes the contributions of the thesis. Finally, an outline of the thesis is provided in Section 1.8.

1.1 The research domain

The Internet, from its birth keeps on the tradition of providing different communication and information sharing services. OSNs represent a recent type of communication and socializing platform [1], which is welcomed by the Internet users and has grown more than two billion users according to Wikipedia's list of prominent social networking websites [2]. Unlike the traditional web which revolves around information, documents, and web items, the concept of OSN revolves around individuals, their connections and common interest-based communities. These online communities share or refer (provide links to other web resources) the traditional Internet resources with each other. An OSN consists of a virtual social graph where users are nodes who are connected with each other through a relationship, which forms the edges of the social graph. According to the Antonio et al. [3], OSN services for an individual are: (1) to create a public or semi public profile where he or she shares personal information such as name, contact, interests etc. (2) to establish a social circle of friends for information sharing and communication (3) to view and traverse friends' profiles and private information (4) to carry out real time and non-real time communication with friends in the form of comments, private messaging, chatting, picture tagging etc. and (5) to use a lot of third party applications that range from simple poking to gaming, advance communication, virtual gifts, event management and so on.

OSN based interactions and social activities have increased privacy concerns because various intruders try to harvest OSN users' data with both positive (for personalized friend, product or event recommendation) and negative intentions. OSN users are unable to deal with these kinds of privacy attacks due to several reasons. In reality, a lot of OSN users are unaware of these privacy breaches and vulnerabilities. Secondly, OSNs mostly provide manual security settings in order to tackle these intrusions which are hard to use. Furthermore, Individuals can also face privacy threats from their own social networks because their network consists of numerous un-trusted and even unknown friends. These internal threats further lead to other serious attacks such as identity theft, profile porting, defaming, blackmailing etc. Unfortunately, users are unable to figure out the malicious members in their social network because OSNs do not provide any manual or automatic mechanism that can be used to differentiate between friends.

Banks et al. have credit to initiate a research in this direction of identifying friendship intensity by introducing the "interaction count" method [4]. Interaction

count may not prove to be a best indicator to predict friendship intensity. Since, individuals do not always prefer to communicate online with their best friends due to several reasons i.e. their context, activity level and interaction habits. In this study, we have indentified metrics and developed ways to integrate these metrics with the interaction count method. Furthermore, this study also proposes data mining framework for friendship levels calculation.

1.2 Aims and objectives

First of all, this study investigates privacy issues, users concerns as well as their expectations regarding privacy control settings provided by OSNs. Later on, this thesis explores the techniques for predicting friendship levels, a basic ingredient to identify potential privacy threats. Finally, this study suggests a framework that utilizes friendship intensity information and tries to cover the most of identified privacy issues.

The fundamental goal of this research is to develop techniques for estimating friendship levels between an OSN user and his/her friends. This research seeks to utilize data mining techniques to achieve this goal. This study first identifies features and metrics that can be used for the prediction of friendship levels. Secondly, this research tries to use data mining classification techniques to solve this specific problem. This thesis also aims an experimental demonstration of data mining phases in order to predict friendship levels. For that experimentation, the training data is created by observing interaction behavior of Facebook users with their good friends. This training data is used to develop data mining model by using couple of classification algorithms. Finally, the performance of these algorithms is evaluated through various statistical techniques.

1.3 Research questions

The main research question of this thesis can be articulated as:

"How can we automatically predict friendship levels in OSNs based on usage or interaction data?"

Our main question is to explore metrics and measures that can be helpful in calculating friendship intensity/levels in OSNs. This is a first step to identify potential privacy risks. These privacy vulnerabilities are quite obvious in OSNs because they are mostly comprised with "weak ties" (more discussion is available in Chapter 2). This study tries to answer following research questions in order to pursue the above problem area which also introduces main application area of this initial study.

RQ 1: Which are the users' concerns regarding their private data as well as their expectations apropos the measures adopted by OSNs and the state-of-art research in this area?

RQ 2: How can we automatically predict friendship levels/intensity from OSNs interaction data?

RQ 2.1: What are the factors that can be used to determine friendship intensity**?**

RQ 3: How can we use predicted friendship intensity to decrease the privacy threats?

It is noted that we will mostly use the terms friendship intensity and levels interchangeably with same meaning. Specifically, friendship levels and intensity can be differentiated where levels may refer to the category of the friendship such as very good friend, good friend or an average friend. Whereas, intensity could be a numerical value describing the strength of the relationship i.e. 70% close friends. In general, we

are concerned about both kind of predictions but we will only consider friendship levels in the experimentation just to abridge the concept.

1.4 Research methodology

A research methodology provides a strategy or an approach to achieve research goals. The selection and proper execution of a suitable methodology is much important to maintain quality and to acquire good results in the research. This thesis utilizes couple of empirical methods such as the survey and experimentation to address the problem domain [5]. The course of research is portrayed in Figure 1.1. This figure briefly describes the applied methodologies and specific RQs which are addressed during each methodological phase.

In the first phase, we perform literature review to cover the background knowledge. A critical appraisal and coverage of this literature can be found throughout the text. This extensive study covers the following main issues:

1) Existing threats and attacks on user private data
2) The state-of-art countermeasures against these attacks
3) Users' expectations and concerns regarding OSN's countermeasures

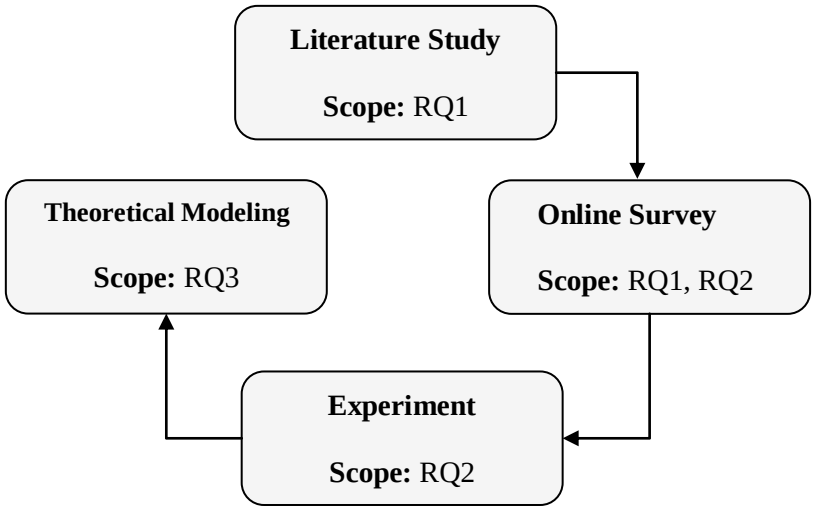

Figure 1.1: Research process summary

After performing the background study, the course of research progresses on the next level where an online survey is conducted in order to answer some part of RQ1 and RQ2. This survey serves dual purposes of the study; identification of users' privacy concerns and their interaction behavior with good friends. The later inquiry provides a basis to the experiment, the other empirical method which is used to answer RQ2. Table 1.1 provides a brief overview of research methods applied, data collection, data type and context.

Table 1.1: Overview of the research methods applied

	Method	Data Collection	Data Type	Context
RQ1	Case Study, Survey	Online	Quantitative	Facebook, Academia
RQ2	Experiment	Experiment execution	Quantitative	Academia

Finally, this study devises a theoretical framework based on the empirical results in the process to answer the RQ3. This framework utilizes calculated intensity/levels for privacy preservation. Throughout this study, we take Facebook as a case study and,

the details of this qualitative method [6] as well as other research methods, are covered below:

1.4.1 Case study

The case study of the most popular OSN website is considered to analyze users' privacy issues and their interaction behavior. There is couple of reasons to choose Facebook as a case study. First of all, Facebook offers comparatively better and comprehensive privacy control mechanism [7]. In their privacy control, Facebook provides numerous features to its users for managing the access of their data and personal information. Furthermore, it is also observed that Facebook users feel more confident in sharing their personal information than other OSNs such as MySpace [7]. Therefore, Facebook could serve as a good case in order to investigate users' privacy concerns.

This study acquires Facebook users' views for analyzing their privacy concerns, their privacy preserving habits, and their satisfaction level regarding Facebook privacy settings. To obtain real observations of Facebook users, an online survey is conducted. This survey is main methodology to cover numerous aspects related to RQ1 and RQ2. The details and motivations of this survey are covered below:

1.4.2 Online survey

An online survey is one of the two major empirical methods that are applied in this study. An Internet based surveying website is used to conduct this survey. In addition to the closed ended survey questionnaire, a Facebook group and email account is established to obtain views and comments of the users. Besides efficient data collection and analysis, the other main motivation of using the online survey method, is to promote non-privacy-violating research methods in OSNs.

The objectives of this survey are: 1) to explore users' concerns about their privacy on Facebook, 2) to find the extent at which users like to expose themselves on Facebook, 3) to analyze users' familiarity with the Facebook privacy measures, 4) to identify users' satisfaction level against privacy measures provided by Facebook, 5) to track users' interaction habits with their good friends and, 6) to find out the interaction types that are commonly used in Facebook. We have developed a close ended survey questionnaire to achieve these objectives; the further details of this questionnaire and survey are available at Appendix 7.2

The selection of survey sample is a critical step which defines overall validity of the results. This research has decided to conduct this survey at School of Computer Science in Blekinge Institute of Technology (BTH), Sweden. The reasons to select this sample are following:

- The nature of our sample is multinational, multicultural and multiethnic, because BTH has student representation of many countries. These characteristics of sample has provided us different/common preferences over diverse aspects of privacy
- Most of the participants in our sample are computer science or software engineering students, who are proficient in computer usage. How proficient are they when it comes to use Facebook privacy settings?
- We observed that a lot of individuals in our sample are frequent Facebook users. They reason is that they are far away from their homes where most of their social connections dwell and Facebook provides excellent platform to keep in touch with them.
- In OSNs, the most of prior study is about teenagers and their apathy towards privacy [8-10]. From our sample, we try to investigate privacy concerns of a little different age group. The individuals in the survey sample, are mostly the graduate students who are little mature than teenagers.

1.4.3 Experimentation

This study also performs an experiment to demonstrate the practical implementation of data mining framework. The main purpose of this experiment method is to investigate whether it is possible in general, to distinguish among different friendship levels by using data mining techniques, or not. Furthermore, this experiment also provides proof-of-the-concept implementation. This study has utilized Weka [11] workbench to perform experimental procedure.

The overall experiment process is portrayed in Figure 1.2. The first step is the selection of attributes which always keep their importance in building efficient decision models. The empirical justification of attribute selection is sought through the survey. We have selected five interaction based and one non-interaction based attributes. Although, we have identified two non-interaction based attributes; profile visits and mutual friends however, only profile visits is used in the experimentation. We have found strong empirical support for this metric in the survey where more than 80% of the respondents like to visit the profile of their good friends. The construction of training data is the second step in the experiment process. This study generates training data artificially by using various indicators that are inquired in the survey. Furthermore, the training set is generated randomly by setting threshold values for various selected attributes. The random process of training data creation reduces the factor of human bias. Training data consists of 404 unique instances which covers each level of the friendship.

After the creation of training data, the experimentation process proceeds to the algorithm selection phase. J48 and Naïve Bays, are selected to create data mining prediction model [11]. Both algorithms perform classification tasks but their way of handling the problem, is fundamentally different from each other. J48 is a decision tree based algorithm which constructs a tree to perform classification decision. On the other hand, Naïve Bayes calculates posterior and prior probabilities for prediction of a friendship level [11]. At the end of experiment, relative performance of these algorithms is evaluated by using statistical techniques such as Cross Validation (CV) [11].

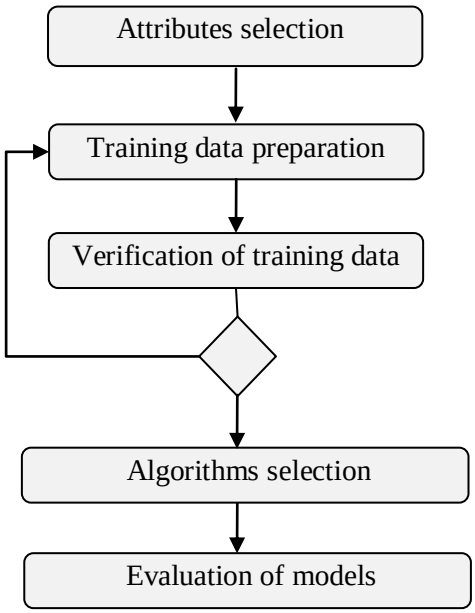

Figure 1.2: The experiment process

1.5 Validity threats

Internal validity, statistical conclusion validity, construct validity and external validity threats are identified for empirical methods by Creswell [6]. These validity issues are subjected to be present with the empirical methods and data collection procedures. The first three validity issues (internal, conclusion and construct) are related to the process of effectively answering the research questions. The external validity threat is about the generalization ability of the research. It concerns, how effectively the results of the study can be applied in a scenario other than experiment. We have applied two empirical methods, online survey and experimentation, to answer RQ1 and RQ2.

In case of online survey, external validity threat may arise because the survey is based on the observations from a small sample of overall population. This study has tried to reduce this threat by taking subjects from every group, such as demographic, age, sex divides. Some of the survey results may have generalization issues because the responses are quite biased on some aspects such as nationality and age. More than 50% of the respondents belong to single nationality. As far as the age of the respondents is concerned, more than 80% of the subjects are between 20 to 40 years. Therefore, the results of the survey are quite applicable for this age groups and nationality. In our analysis of results we did not consider the discussions specifically based on age and nationality. However, we expect to keep this limitation of results while generalizing it on other similar scenarios. Furthermore, the closed questionnaire of survey may reduce the responses to some certain context. This study has struggled to include every possible options/scenarios as the answers of survey questions to reduce this validity issue. In addition to this, a Facebook group and an email account is created to further acquire respondents' views that are expressible in the survey.

We can also face validity issues regarding the conclusions, process and setting of the experimentation. The first type of threat that may arise is internal validity threat which is related to the procedure of the method. We have tried to reduce this validity issue by utilizing the already implemented and tested methods for experimentation. Moreover, another major validity risk related to the creation of artificial training data for training of the data mining model. To mitigate this threat, the artificial training data is created randomly by using the survey results.

1.6 Related work

OSN is a diverse subject who has roots in multiple areas. We try to cover various aspects of OSNs throughout this text and the related literature can be classified into following three categories:

1.6.1 Online Social Networks

People do not live alone but they are bounded with other people through numerous invisible connections that form families, clans, societies, countries and the world. These connections also referred as "social structures" or ties. A lot of research is done in this subject under the areas such as Sociology, Psychology and Anthropology. The idea of OSNs is inspired by the work of social psychologist, Milgarm [12]. In his research, he conducted an experiment to prove that an arbitrary person needs maximum six connections (acquaintances) to reach any other randomly chosen person. This phenomenon is also famous with the names of "six degree separation" and "small world" [13]. Other influential work is done by Granovetter [14] where he has argued to classify social connections into "strong" and "weak" ties. Furthermore, he has highlighted the strength of weak ties by claiming that more weak ties are important to get new information and a variety of other benefits. According to him, cliques (strong connections) could fall into homophillic tendency (such as homogenous traits and opinions because of usual like-mindedness in such groups). The range of other work on OSNs, covering their representation, benefits and related issues, can be found in

[15-18]. An important research direction in both real as well as online social networks, is analyzing their growth, connections and interactions patterns. This area is known as social network analysis, and covered in detail by Wesserman [19]. Another related work in formulating and explaining different phases of social network analysis, is conducted by Laura et al [16]. With the emergence of OSNs, scientists start studying the behavior of this computer mediated social networks. Most of the times, they try to relate OSNs with real life social networks and they found extraordinary similarities. Several studies are conducted to prove social phenomena such as power-law, scale-free growth, preferential attachment, small world [20-23]. A large scale study on four OSNs, is conducted by Mislove et al. [21], where they observed that social networks contain densely connected core of high degree nodes which is connected by small group of low-degree nodes.

OSN websites brought recent trend on the Internet by providing a platform for developing and preserving social connections. The most comprehensive discussion on OSN websites as well as their history, types and other issues, is conducted by Boyd et al. [1]. The authors have covered many aspects of social networking sites in their work. Furthermore, Wikipedia[1] resources on social network sites also provides comprehensive discussion, news and facts about OSN websites [2, 13].

1.6.2 Privacy and security of OSNs

Apart from the benefits, OSNs give rise to privacy and security threats over the Internet, more severely than before. Lots of studies are conducted to understand peculiar human nature about the privacy of their data. Discussions about contradictory privacy preferences among human and its importance with respect to modern information age can be found in [24] and [25]. Boyd [25] highlights human aspects related to "privacy" and "publicity". According to her individuals publicize their data cautiously to gain some benefits or instant fame.

The research conducted by Gross et al [26] and Krishnamurthy [27], are two influential studies in OSN privacy and security. These authors highlight many privacy issues of OSN sites. Gross in his study draw attention to the factors that compel users to reveal their most private data. Krishnamurthy has the credit of characterizing different aspects of privacy. He has conducted a study of many OSN sites in order to analyze the privacy controls provided by these OSNs. Furthermore, he also argues to divide user data into small chunks for providing more efficient access control. Additionally, he further suggests that OSNs should only provide required data to the third party application and games. Finally, in other studies, the researchers cover security and privacy threats [28, 29] to the social network data at different levels [3] and their solutions [30-32].

1.6.3 Calculating friendship intensity through data mining

We did not find any research which applies data mining algorithms to calculate friendship strength in OSNs. Therefore, this research is the first stride in this direction.

As far as calculation of friendship intensity is concerned, we found an interesting ongoing study, that is conducted by Banks et al. [4]. The authors have introduced interaction count method in which they take different types of interactions and count them in order to calculate friendship strength. In addition to provide a novel intensity calculation method, they also suggest a framework that utilizes calculated friendship intensity for better privacy control in OSNs.

1.7 Contributions

This thesis is concerned with the overall improvement of user privacy in OSNs. Followings contributions are made while answering each of the research questions that are stated in Section 1.3:

[1] www.wikipedia.com

1.7.1 OSN and privacy

Besides covering privacy issues of OSNs, this study also provides a comprehensive overview of almost all related aspects. First of all, this study has devised a classification for OSN websites into dedicated OSN and multipurpose OSN. The purpose of this categorization is to provide a formal division of various social networking sites available on Internet. Later on, this study also categorizes the solutions against the various privacy attacks into technical and precautionary measures. In addition to this, two types of decentralized privacy preserving solutions are found in the literature. These solutions are differentiated as multi-OSN and single-OSN decentralized networks in this work.

This study has identified various characteristics of the users related to their privacy. These aspects correlate users' behavior with their expectation while preserving their private data. Furthermore, a variety of users' views are gathered in this study regarding their private data, their satisfaction over privacy controls, and privacy interference.

1.7.2 Friendship levels prediction

Friendship intensity calculation is the major target of this research activity. This thesis claim following contributions while performing this dimension of research:

- The main contribution of this research is the suggestion to use data mining for friendship intensity/level prediction. Besides recommending this machine learning approach, this study also provides a framework that describes the ways to use data mining techniques for that purpose. No doubt, this framework is mostly inferred from the traditional data mining process but we generalize this basic process into OSN context. This study also introduces a method to structure the training data for data mining algorithms.
- Banks et al. [4] has suggested interaction count method for calculating friendship intensity. In extension to their research, our study has identified three major issues such as interaction habits of individuals, their activity levels and their context. This study also found that the relative importance of interaction type vary from user to user. Furthermore, various ways are discovered to integrate these features with interaction count method to improve its performance.
- In addition to the enhancements that we have suggested in interaction count method, various other metrics are also suggested in this work to calculate the strength of a relationship. These factors include commons friends, profile visits, interaction content and context.
- Finally, this study has devised a framework that utilizes the friendship levels information in order to achieve the semi-automatic privacy control mechanism in OSNs.

1.8 Thesis outline

The rest of the thesis is organized as follows:

In Chapter 2, the background of the problem area is introduced quite comprehensively and thoroughly. This chapter begins with couple of sociological theories that are also considered as basis for OSNs. Afterwards, the area of OSNs is covered with many related sub-areas such as social network analysis, types of OSNs and analysis of OSNs. Later on, this chapter elaborates the structure, facilities, and history of OSN websites. After covering the introduction to the subject, this chapter moves on the privacy attacks that are faced at different levels of an OSN. In the end, this chapter covers various techniques that are proposed in the literature to safeguard user data in OSNs.

Numerous aspects of user privacy in OSNs are identified in Chapter 3, which answers most parts of RQ1. This chapter starts with the discussion of privacy in real life social networks and its correlation with the computer mediated social network.

This conversation strives to relate users' privacy preserving habits in offline setups with their online equivalents. Later on, this chapter moves on the topic of online survey that is used for gathering views and concerns of Facebook users. This section starts with the motivations of survey methodology, its design and analysis methods. After that, the presentation and analysis of survey results are provided in detail. This section presents and analyzes users' views on various issues such as potential privacy threats; OSNs provided privacy settings and concerns over governmental interference. Finally, the summary and discussion of survey results, is provided at the end of the chapter.

Chapter 4 answers the research issues raised in RQ2 and RQ3. This chapter begins with the justifications and potential benefits of calculating friendship intensity. The next section introduces several interactions based methods such as interaction type, interaction count, interaction content and context. This part of study also identifies potential limitations with the interaction count method and ways to revamp this method. After that, this chapter discusses mutual friends and profile visits as two non-interaction based techniques for calculating friendship intensity. Additionally, this part of chapter also presents the empirical justification of these metrics. The data mining framework for calculating friendship intensity/levels with their benefits of utilizing this technique is covered, afterwards. This chapter also covers the process of experiment on the two classification algorithms to demonstrate the practical implementation of the concept. In the end, this chapter introduces a framework that utilizes friendship levels information to improve the privacy mechanism of OSNs.

2 ONLINE SOCIAL NETWORKS AND PRIVACY

OSNs are conglomeration of almost every communication and collaboration technology used on the Internet. Moreover, OSNs are utilized to preserve, animate and enhance the social connections of individuals with other individual or group of people. The purpose of this chapter is dual. The first part presents an overview of the social network, OSN, Social Network Analysis (SNA), OSN websites and related concepts whereas the subsequent part covers OSN privacy issues, threats and state-of-the-art counter solutions. This chapter is organized as: In Section 2.1, the couple of social network theories are introduced that led to the development of OSN applications. The idea of social network with related concepts i.e. SNA, the area which is revitalized with emergence of OSN, is discussed in Section 2.2. A simple overview of OSN websites is provided in Section 2.3. In the second part, user privacy related issues and privacy attacks on OSNs are covered in Section 2.4 and Section 2.5, respectively. Finally, some of current counter solutions to these attacks are provided in Section 2.6.

2.1 Introduction

Computer network also serves as social network when they are used to connect individual and organizations [16]. The idea of social network is as old as human started living together but its benefits and effects were perceived not so long ago. The evolution and effects of social networks have been mostly studied in the areas such as Sociology, Anthropology and Social Psychology. However, with the advent of various communication technologies, the channels of interaction and socialization have been changed, considerably that not only break the communication barriers i.e. distance, time but also switched this area very much multi-disciplinary.

Almost, all theories of social network mostly focus on the importance of connections between people not individuals alone. According to Karol Mark's, " society is not merely an aggregate of individual; it is the sum of the relations in which these individual stand to one another" [15]. Modern social network theory is based on the experiment conducted by Stanley Milgram [12] in 1967. Milgram asked several subjects (people) to forward a letter to his associate at Boston by passing it to the people on the basis of first name acquaintance. The purpose was to pass this letter through fewest numbers of "hops", intermediate people; the average was 5.5 in that experiment. Milgram concluded that any two randomly selected people in USA are at most 6 levels away from each-other. Apart from the exact number edge distance, this experiment proves that average distance between two individuals is not very high. Figure 2.1 visualizes the process of this historical experiment.

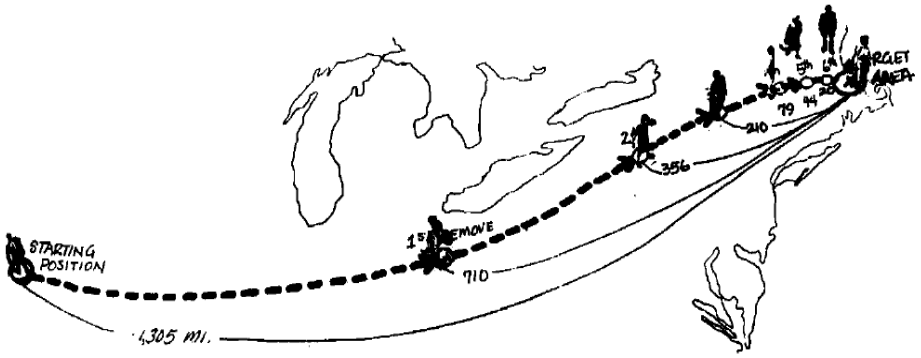

Figure 2.1: Small world phenomena by Milgram[12]

Other significant work in this area, is done by Granovetter [14]. In his work, he classified social connections (friends) of people into 'strong' and 'weak' ties. He

further emphasizes on the importance of weak ties in his work. It is observed later that networks with lots of loose or weak connections can be more valuable to its members than the tighter social networks. Furthermore, social networks with many weak ties are also referred as open networks. Open networks are more likely to introduce new ideas and opportunities to their members than the closed networks with many superfluous ties. The net benefit that a social network can bring to an individual or organization, is also known as Social Capital (SC) [15].

The Internet turned abstract social connections more practicable by making it digital over physical wires of computer network. Moreover, Internet offers a variety of interaction and communication facilities to the individuals for maintaining these connections. OSN sites, i.e. Facebook[2], MySpace[3], and Twitter[4], provide facilities to explicitly declare and enhance social connections. In other words, it provides an opportunity to flourish "weak social ties" of individuals. These OSN sites also offer almost all interaction and socializing facilities in a single place. Besides communication, it is also observed that OSN sites increase SC or social benefit more than any other Internet service [17]. The recent location based mobile feature which notify users to their nearby friends, reduces the gap between online and offline communication [17]. OSN sites also support in coordinating and mobilizing social actions performed by organizations, political figures, regional associations [17].

2.2 Online Social Networks

In a social network , group of people are connected with each other and with other groups through a relationship [15]. The main idea of social network constitutes a large social graph where individuals or organizations are nodes that are tied with other nodes through a relationship or tie. These ties or interdependencies emulate many forms such as friendship, kinship, co-worker, co-authorship, information exchange etc. Social networks reflect the pattern in which these individuals are related to each other. In computer mediated communication (CMC), much of the research is focused on how people interface with computers, how individual interacts using computers and how groups of people cooperate [16].

Mathematical tools such as matrices and graphs are used to depict the social network phenomena [18]. The nodes of the graph are used to represent the individuals where edges describe the existence, intensity (weighted edges) and direction of the relationship. The graphs of this type are known as Sociograms [18]**.** Figure 2.1 exemplifies sociogram of an email based social network which consists of almost six thousand nodes and more than 100 thousand edges.

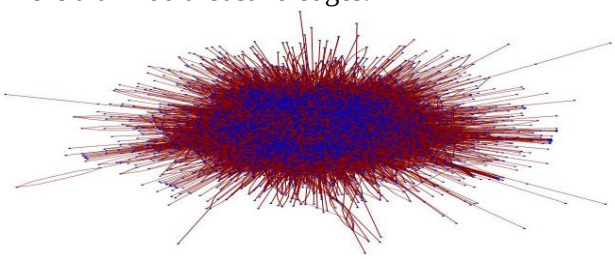

Figure 2.2: Sociogram of email network [33]

Sociograms are good tool for SNA but they may not be much comprehensible when many individuals are part of the social network as evident in Figure 2.2. Matrices are alternative tools to represent the social network relationship. Table 2.1 shows matrix representation of friendship relationship among three individuals. In this table, binary one indicates existence of relationship whereas zero refers no friendship.

[2] www.facebook.com
[3] www.myspace.com
[4] www.twitter.com

Furthermore, numerical values can be used in spite of binary values to show the strength of the relationship. The matrices representing the social relationship, are also called Sociomatrices [19]**.**

Table 2.1: Matrix representation of Social Network

	George	Michael	Pam
George	-	1	1
Michael	1	-	0
Pam	0	1	-

Individuals establish relationships in diverse context and circumstances. Social networks can be classified according to the context of the relationship such as business and social [15]. The relationship instances of business context could be colleague, co-worker, co-author etc. On the other hand, social context might include relations of type i.e. friendship, relative, neighbor [15]. These contextual relations are not mutually exclusive imply that; individuals could be both friend as well as co-worker. Social networks are also classified on the basis of technology adoption where personally social and device supported social networks (DSSN) [15], are two such divisions which are portrayed in Figure 2.2. Both types of networks carry pros and cons where DSSN do not consider non verbal communication i.e. facial expression, voice tone during interaction but they support to evolve social relations without any boundaries of countries, region etc. Nonetheless, the continuous development of communication technologies is reducing the gap between online and offline interaction. Nowadays, many of the features which are part of offline conversations can be seen in online communications.

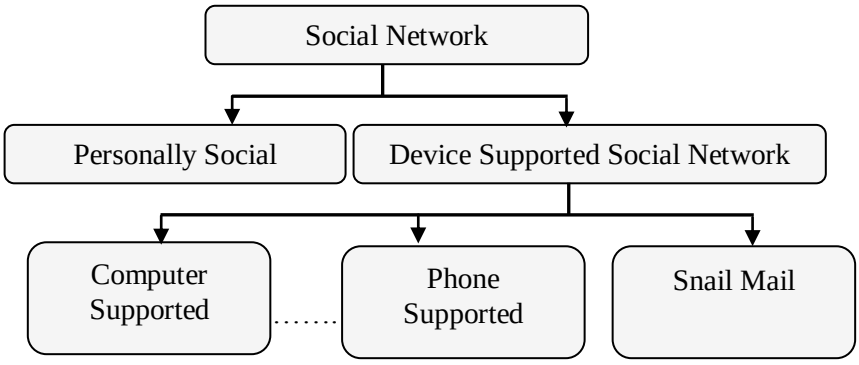

Figure 2.3: Social Network Classification [9]

OSNs have great advantage which individuals start reaping just after the invention of computer networks, Internet and the Web. The computers and computer networks has a phenomenal role in preserving and enhancing the social networks. But these kind of networks reduces the element of "social presence" which may miss some of verbal (i.e. voice tone) and non-verbal (i.e. facial expression) signals [34].

OSN provides a variety of communication methods to its members which are classified as real-time and non real-time. In real time communication, the presence of each party is required to perform real time or synchronous interaction i.e. audio and video chats. On the contrary, the presence of an individual is not required in case of non-real time communication which is also known as asynchronous communication. The interactions of these types include email, comments etc. The migration of Web into interactive Web gives birth to a variety of non-real time communication methods where users can write their feedback about every Internet resource in form of comments.

2.2.1 Social Network Analysis

The evolution, interaction patterns and growth of social network contain a lot of interesting realities that can be examined by performing a systematic study of social networks. This systematic study is termed as Social Network Analysis (SNA). Although, SNA is performed from last 50 years in social science such as Sociology, Social Psychology and Anthropology but with the emergence of OSNs, this research has been revived, considerably. On one hand, OSNs provide a laboratory to experiment many of social theories such as Small World Phenomena, Weak Ties. On the other hand, OSNs are also revealed several social aspects regarding individuals, groups, organizations and nations which were never observed before. The SNA analyst tries to cover the network of relations as fully as possible to analyze the flow of information, and to observe what effects these relations leave on individuals and organizations. According to Laura et al. [16], SNA is performed in following phases:

- Sample Selection
- Data Collection
- Data Analysis using SNA method
- Conclusions

In SNA, first of all the target group selection is performed to identify the patterns that exist in that particular network. This selected group of nodes and connections is called sample or population [16, 18]. Many methods are used to collect information from that sample. These methods include questionnaires, interviews, observations, diaries and through computer monitoring [16]. In OSNs, crawling and monitoring methods are quite common to observe the overall structure of network. To crawl the OSN, an artificial agent or simple software move from one node to the other (from friend to friend) systematically (commonly used method are Breadth First Search (BFS), Depth First Search (DFS) and their variants). The movement of this software agent is then simulated to observe the overall structure of network. In addition to this, recording the patterns of interactions between individuals over a longer period of time, is another commonly used method in computer based social networks. The type of the data which is collected from the sample is also called units of analysis which consists of relations, ties and actors [16].

After the collection of data, the next step is to analyze this data by using methods such as full network , snowball and ego-centric methods [16]. The Full network method provides whole picture of the network but it becomes complex when there are many actors in the network and each possible connection between actors need to be considered [16]. Moreover, a complete list of connections between people as well as their links to external environment is created while performing full network analysis which makes this process quite resource consuming [15, 16].

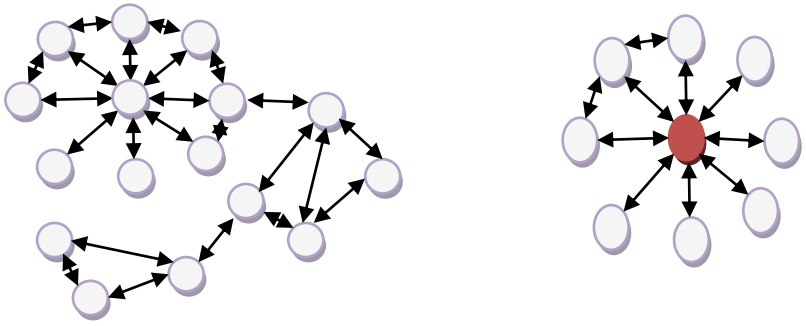

Figure 2.4 (a): Whole Network (Socio-centric) **(b):** Ego-centric

The ego-centric method is quite useful when complete network analysis is not required [15]. In this technique, first the “ego”, starting point of analysis, is identified and then the process is moved to his “alters”; connections that are one step away or friends in most OSNs [15, 18]. This technique has a variant which is called “ego only”,

where analyst only consider first level connections [18]. The Figure 2.5(b) depicts the egocentric social network where red node represents the ego and remaining nodes are alters. If the identified alters in an egocentric network becomes egos to continue the analysis, then this method is referred as Snowball method [18]. Snowball method is also time and resource consuming and one recently developed method which covers the limitation of Snowball method, is referred as Hybrid method [35]. Hybrid method only tries to analyze important alters despite considering all alters, and these alters are nominated by the egos. Finally, conclusion is the last step in SNS where analyst tries to come up with new findings or just prove/disprove some hypothesis.

2.2.1.1 Online Social Network Analysis

A lot of work is carried out in recent years to analyze patterns and growth of OSNs in a manner that how they affect or affected by real life social network. Researchers observe intriguing similarities between OSNs and traditional real life social networks [20]. Barabasi proved that OSN has power-law, scale-free growth and exhibit preferential attachment [20]. More influential research on OSNs is conducted by Mislove et al. [36] and Ravi et al. [22, 23]. Mislove et al. performed a large study of four OSNs; Flickr, YouTube, LiveJournal, and Orkut. They collected the data of 11.3 million users and 328 million links by crawling publically accessible profiles. They proved many of social theories i.e. power-law, small-world, and scalefree properties of OSNs from their findings. They also examined that these networks contain densely connected core of high degree nodes which is connected by small group of low-degree nodes.

OSNs can be viewed as a graph, G= (V, E) where V shows set of nodes i.e. v_1, v_2 …v_n and E represents set of edges that connect these nodes. Two nodes are connected in the graph, if an edge exists between these nodes. Furthermore, if we can reach from a node v_i to v_j by passing through one or more intermediate nodes then the path between v_i and v_j exists in that particular graph. The path between two node in a graph is denoted as $v_i \rightarrow v_j$ which represents sequence of nodes that should be traversed to reach v_j. A graph is strongly connected if for any two nodes there exist a→b and b→a. Moreover, a graph can be sliced into one or more sub-graphs of strongly connected components (SCC) [37]. In SNA, SCCs are used to identify strongly connected sub-groups in the social graph. Some important metrics that are used in OSN analysis are described below:

Size: The size of G is denoted as n = |V|, it represents number of people in the social network.

Density: Density of the graph or SCC shows the ratio between the individuals and their relations (connections). The minimum density of a graph is 1/n (it is the case when graph is a ring) and maximum density is 1.

Diameter: Diameter of a graph shows maximum length between any two nodes. The diameter of a social graph is between 1 and n.

Adjacent matrix: Adjacent matrix is just a matrix representation of the graph or SCC. This is a matrix of size n × n where $a_{i,j}$=1, if $v \rightarrow v_j$ exists and 0 otherwise.

2.3 Social network sites

Social network sites or OSN sites are type of OSNs which have revolutionized the Internet. Unlike other websites where documents are linked with other document, in OSN sites people are linked with other people to form computerized social network. OSN sites has magnetized numerous Internet users in just last five years which also open a window of new research opportunities in many areas such as Sociology, Anthropology, and Computer Science etc. There are hundreds of OSN sites with different technological capabilities, supporting a wide range of interests and practices [1]. Many of OSN sites started with the concept of "social networking" means; people will develop new online acquaintances to expand their social network. But this is

mostly not the case with OSN sites where individuals only like to automate their "latent ties"; the people to whom they share real life connections [1]. Therefore, people only like to share among their real life network in the most of large OSN sites.

OSN sites did not observe much excitement in the beginning when first website (sixdegrees.com) of this type was launched in 1997 which only survived for three years and its founder thought that "its ahead of time" [1, 2, 13] .The peak time of OSN sites' emergence and popularity, was from 2002 to 2004 when some of famous OSN sites i.e. Friendster, MySpace, Bebo, LinkedIn and Facebook were launched. OSN sites has started to flourish from 2005 onward and at present, OSN sites are among the top Internet websites in terms of user base and Internet traffic [2]. Table 2.2 presents top five OSN sites, their user base and website rank.

Table 2.2 : Top OSN sites [2, 38]

OSN sites	User Base	Web Rank
Facebook	400,000,000	2
QZone	200,000,000	10
My Space	130,000,000	17
Twitter	75,000,000	12
LinkdIn	60,000,000	29

OSN sites are classified in terms of their use, features and purpose. One classification is internal social network (ISN) and external social network (ESN) [13]. In this division, the former type of social networks comprises closed/private networks within a society, business or organization while the later type is open/public social network which is opened for everyone to create and evolve their interest communities. Most of the large OSN websites are instances of ESN i.e. Facebook, MySpace etc. Besides this categorization, social network websites can also be divided according to their purpose or some particular interest. In this regard, OSN sites are divided into two broad classes: dedicated social network (DSNS) and multipurpose social network websites (MSNS). In DSNS, social networks are developed to perform some specific pursuit or task i.e. dating, picture sharing, video sharing. Livejournal[5], YouTube[6] and Date.com[7] are examples of DSNS. On the other hand, multipurpose OSN sites allow performing almost any activity according to one's own interest. MSNS instances include Facebook, MySpace, and Twitter etc.

2.3.1 Features of OSN sites

Besides a variety of exclusive features, OSN sites also increased the utilization of numerous already available Internet resources and applications. OSN sites provide various features which range from socializing with friends to sharing or recommending external web pages or resources (i.e. hyperlinks, videos, news etc.). These features are more or less same in the top social utility providers. Figure 2.5 depicts the process of using OSN sites from initiation to its continuous utilization.

Like most websites, user registration is the first step in an OSN site. After registration, user is asked to create his profile which comprises various type of information i.e. picture, contact, education, address, interests etc. The user profile consists of user's personal information which could be alluring for potential privacy attacks. In these circumstances, the profile visibility is an important issue which depends on the site's privacy policy and user discretion [1]. Some of the websites allow external users (individual who are not even part of OSN site) and applications (crawlers) to view or extract information from the user profile in their default settings. In Facebook, profile visibility varies at different levels such as friend, friend of friend

[5] http://www.livejournal.com/
[6] http://www.youtube.com
[7] http://www.date.com/

or external user. Friends can view the profiles of their friends but profile visibility for other levels depends on the user's own choice.

After creating his public/private persona, user can create relationships with other members of that particular website. This relationship is mostly labeled as friendship, fan or contact [1]. Friendship is most commonly used relationship in almost all OSN sites which may not depict exact description of the relationship that some of individuals may be bonded in reality. Later on, individuals search or invite their friends to form a user-centered social network. In most of the websites, the friendship relationship is bi-directional which means friendship confirmation is required from both sides. For example, if X sends friendship request to Y then Y's confirmation/acceptance is required to be a part of X's social network. Some websites also allow unidirectional relationship in form of fan or follower [1]. While dealing with user privacy, the visibility of social network (Friend's list) is also a crucial aspect for OSN sites. The majority of the websites permit everyone to view or traverse the friend's list of a particular user, but some also facilitate their users to control the visibility of their friend's list. Moreover, traversing someone's friend's list (or social graph) is most basic activity in SNA.

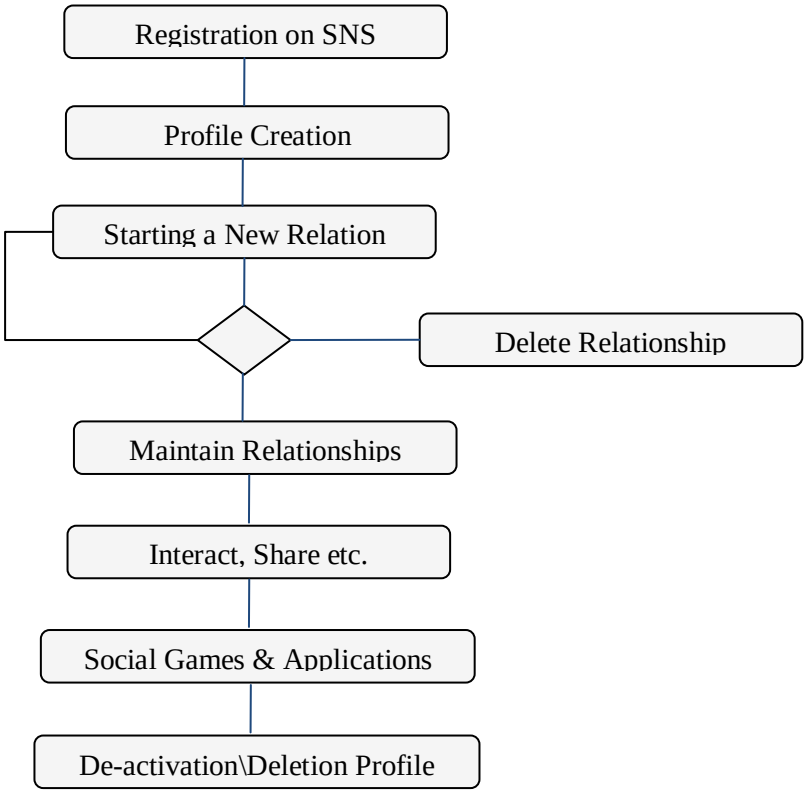

Figure 2.5: Function of OSN sites

After forming the first social connection, users can add more friends in their network by following the same procedure of sending friendship requests. Users can even remove a social link by simply deleting that particular contact. OSN sites provide a variety of interaction and socializing facilities to communicate with their friends and to remain active. These features include private messaging, chatting, and comments etc. Sharing is one of the powerful facilities provided by OSN sites. An OSN site user can share virtually any Internet resource such as pictures, videos, hyperlinks with their social network. This is an important utility in many ways; it increases the viewership and utilization of several other websites and Internet resources. Additionally, it provides kind of social authenticity to some material from very huge information source, the Internet.

Some OSN sites also provide different type of games and applications which are mostly developed by third parties. These games and applications are an important reason behind the success of several OSN sites such as Facebook. Alongside games, OSN sites also provide a variety of applications that serve different purposes i.e. birthday wishes, daily horoscope, visualization of social graph. In addition to these features, some more facilities are available on OSN sites such as picture uploading where users can upload their pictures and create albums. They can even tag these pictures with the names of their friends. Other facilities consist of creating blogs, events, ads, and movements etc.

2.4 OSN and privacy

With the emergence of OSNs, privacy concerns has become the flash-point on the Internet and there exist several dimensions of user privacy such as political policies, the rights of citizens and consumers' protection [24]. Privacy is a complex human characteristic which varies person to person and information to information. Additionally, social factors, education level, age, popularity and wealth affect privacy preferences of individuals [7]. In many circumstances, individuals want their information should only be known to a small group of close friends, not to the outsiders and on the other hand, they want to reveal their information only to strangers but not to their close friends [26].

The user privacy on Internet deals with user's ability to control: 1) what information he wants to reveal on the Internet and 2) who can access and use this information. Many people think that Internet is "public" which ensures no privacy and if someone is sensitive about his privacy, he should not be there. Recently, Google's CEO, Eric Schmidt, is asked in a TV interview[8], should Google's users treat the search engine as a "trusted friend."? His response was;

"Judgment matters, If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place, but if you really need that kind of privacy, the reality is that search engines — including Google — do retain this information for some time… it's important, for example, that we are all subject in the United States to the Patriot Act and it is possible that all that information could be made available to the authorities."

A strong public disagreement over his views can be found in the form of many comments which he received after this statement[9]. Perhaps, privacy is still one of the dearest human aspects which people do not want to compromise at any cost. One of the notable Internet communication expert, Dana Boyd, also criticizes by arguing that people are still very much sensitive about their personal data, even in this "public Internet", just like they are careful in real world which is also public [25]. She further uttered, the understanding of privacy is little different in this highly public and corporate environment. In the following section, we try to analyze those aspects which induce the users to reveal their most private information in spite of having limited control facilities.

2.4.1 Information revelation and user control

The OSN user reveals his/her private information with varying purposes such as to gain some benefits, to gain popularity, and to remain in touch. These factors are further elaborated next:

- Some people do cost-benefit analysis while revealing their private information. It means, for some users the perceived benefits of information revelation are higher than the costs of privacy invasion [25].
- As it is mentioned before, OSNs are generally the digital representation of real life connections. In this situation, peer pressure is another factor that enforces

[8] http://www.youtube.com/watch?v=A6e7wfDHzew, retrieved on April, 2010

individuals to reveal their personal information and most inner thoughts in order to get appreciations or consolations.

- It is observed that user's attitude towards online privacy is much re-active; they mostly respond after facing some real privacy breach or threat [4]. This attitude gives rise to a relaxing or herbing attitude towards privacy [26].
- Members' faith on some specific OSN site, is another important reason that gives confidence to an individual to disclose his private information [26]. For instance, Facebook users feel more confident in sharing personal information than MySpace users [7].

OSN user reveals his private information in different patterns, styles and formats. To analyze the access and controllability, Krishanmurthy et al. divided user profile information into five categories; thumbnail, greater profile, list of friends, user generated contents and comments [27]. These categories are ordered with increasing levels of details regarding the private data. OSN sites provide different facilities in terms of controlling the access of these fragments of private information. Most OSN websites reveal almost full users' information in their default privacy settings. Therefore, default privacy control of OSN sites such as Facebook and MySpace, is heavily criticized of being public. It is explored in previous studies that users do not change their default privacy settings. For example, a study of Twitter revealed that almost 99% of users retain their default privacy setting [27]. Furthermore, the information even in a public setting is "private" by default and made "public" with effort [25]. For instance, in a public place our discussion with a fellow remains private until we raise our voice to make that information public.

For external/third party games and applications, several OSN websites such as Facebook provide binary type of control; no access or complete access of user's information. Some researchers argue to develop a mechanism to grant only selective information access to these third party domains [7, 27].

2.4.2 Who wants my private data?

There are numerous potential consumers of users' personal information in OSNs who are mostly classified according to their intention or purpose and that could be either "good" or "bad". The initial user of this private information is; of course, the hosting site. These sites can use this information or extend this information, knowingly or un-knowingly [26, 28]. In fact, the market value of several OSN sites depends on the magnitude of users' personal data, which they hold. One of the most data affluent OSN site, Facebook's net worth is more than five billion US dollars which makes it one of top OSN site in term of market value [39].

Other consumers of various types of users' information are the business and corporate world. Normally, these people require mass level user data to perform market research and analyze new trends to promote their products. For that purpose, they collect a variety of user data such as their activities, interests, emails, address information for targeted marketing. These intruders apply different fair and unfair means to collect such information on the Internet. In another scenario, when the customers transact through the websites of these organizations then they mostly collect their behavioral information without informing them. To achieve that, they use different ways such as usage or behavior mining as well as web mining techniques [40]. Although, these people often claim that the purpose of this privacy intrusion is to improve their services but they have other rationales as well [28]. Many of well known companies such as America Online (AOL) and Intel are criticized for selling customers' phone numbers and on making a chip to identify the user, respectively [28].

Government agencies, researchers and policy maker are other major consumers of user's personal data. Data consumers of this class often claim that their purpose is to protect the OSN users and other people from any kind of potential harm. Especially, government agencies often justify their right on user data through law. The other major users are OSN researchers, who often crawl or monitor OSNs with an apparent

purpose to identify communication patterns, network growth patterns and potential risks etc. Moreover, these researchers also publish their findings in journals, conference proceedings, books etc. which is also violation the user privacy. The access and use of such information by these entities, is might be justifiable but, the question is; do we have norms and codes of conduct for information access by these information recipients?

The information consumers discussed in the preceding text, do not have clear intentions to harm the user, at least not to a good user. There are also information consumers who access users' information with wrong intentions. The purpose of these intruders is to physically or mentally harm the information owner.

2.5 Privacy risks in OSNs

There are many privacy risks which are hovering over the user data. Most of these threats can be characterized by different ways i.e. purpose of the attack and according to the social network layers. We adopt the notion of different OSN levels/layers described by Cutillo et al [3], in order to cover OSN privacy risks. According to these authors, social networks can be divided into three levels as depicted in Figure 2.6:

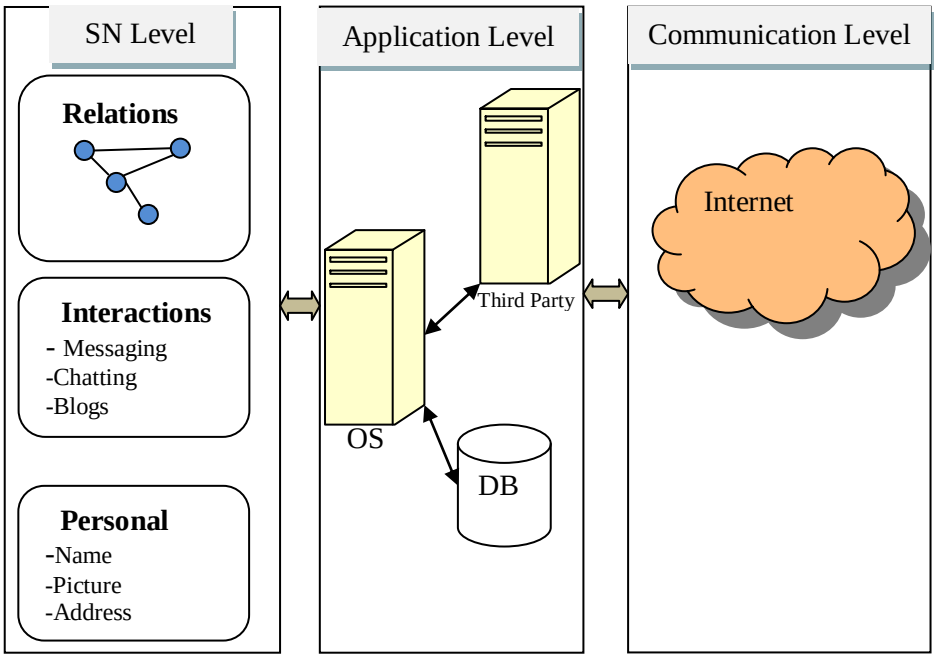

Figure 2.6: Levels of OSN

- **Social Network Level**: This level is digital blue print of user's physical persona, his social connections and his online social activities. Information generated at this level fall into many categories. In general, this level contains user's personal information that comprises his name, phone no, interests, hobbies, political views, sexual orientation etc. and communication traces of the user with his friends. Moreover, his friends' data, their names and pictures are also placed at this layer of OSN. This level also provides many services for communication, socializing, entertainment, sharing, blogging etc. which are used by OSN users to further populate OSN sites with their private data.
- **Application Services Level**: This layer is under the control of social network service providers to perform their activities such as social services, security, storage and management. The services at this level can be divided into two broad categories; social networking services and third party services. To perform social networking services several application, backup and database servers are

utilized by OSNs. Besides OSNs, third party application service providers/domains are other major participant at this level. These vendors provide various applications and games to the end users of OSNs. Most of these applications utilize user's social graph and his personal information to perform their activities. Third party advertisers and data aggregators also perform their activities at this stage.

- **Communication Level:** This level consists of traditional computer networks and communication channels which are used to transmit these social interactions. Almost all OSN sites use the Internet to perform their activities at this level.

Social networks face multiple privacy threats at these levels. Based on the above architecture an attacker can be: 1) A wicked user at social network level, 2) A service provider (mostly third party) at application service level or 3) An eavesdropper; who has access to the communication infrastructure, mostly, through Internet service provider (ISP) [3]. From here onward, we will explore different threats and vulnerabilities at these levels of OSNs. The threats at network level are not covered because they are out of the scope.

2.5.1 Privacy lapses at the social network level

In this section, we look at the privacy threats which are assailed at social network level. Major target for any OSN is to protect member's identity at this level. Attack to the member's identity, is sometimes referred as identity theft. In this attack, the malicious member behaves like the legitimate user by acquiring his credentials. After losing his identity, the user is vulnerable to many attacks. An attack based on this information is known as plain impersonation in which malicious user creates clone or fake profile of the valid user and send requests to establish friendship or defame him. Another related attack is profile porting, in which victim's fake profile is created in some other OSN. In profile porting attack, the attacker makes sure that legitimate user is not present in this new OSN site before creating his profile.

User profile is provide basis for many profiling attacks where user's data provide initial information to the attacker that can be used to guess the values of other important information. Besides names and pictures, users also provide date of birth, income information, interests etc. and this type of secondary user data is harvested to guess other important data. For example, first six digits of Swedish personal number (PN) can be identified from the date of birth of an individual.

Above mentioned impersonation attacks happen because of the fact that not even a single OSN site ensure that the profile is attached with a single person [3]. Faked profiles are common problem which some of OSN sites tried to deal with, i.e. MySpace, but are unable to completely remove them. This phenomenon leads to many of privacy attacks i.e. Sybil attacks, defamation and ballot stuffing. Sybil attacks are most common on peer to peer networks, where multiple fake identities of a user are created by the malicious attacker [28]. If the purpose of Sybil attacks is to forge the reputation of the users then it is also referred as defamation and ballot stuffing [3].

A very common attack at this level is phishing and spear phishing attacks [29]. In phishing attack, mass level deceptive emails (this attack is executed through email, mostly) are send to get some sensitive information. On the other hand, spear phishing is targeted version of phishing attacks where only high profile people are targeted. OSN sites are mostly damaged by this type of attack because of the easy availability of identity information and it took over MySpace at the end of 2006[9]. Furthermore, experiments at different setting show that the success rate of phishing attacks on OSNs, is over 70% and that is why, the phishing attacks on social networks are specifically referred as social phishing [41].

[9] http://en.wikipedia.org/wiki/Phishing

2.5.2 Privacy threats at the application service level

Application service level is the place where social networks functional logic is performed. Most common attack at this level is Denial of service (DoS) attack or distributed denial of service attack (DDoS) [3, 28]. The intention of these attacks is to make some web resource or information unavailable by sending many requests, simultaneously. DDoS mostly serve two purposes of attacker; disruption of information and discontinuation of communication [3]. In addition to this, a malicious attacker, who has privileges to access OSN sites' resources can also intrude into user privacy by performing communication tracking, where he can identify, who is talking who [3].

Major risk at this level is from third party application domains [27]. These domains offer many application and games to the OSN user for a variety of purposes. Most of these applications demand user's personal information and his social connection information to execute their logic. Normally, users allow the access of their private data, which works on the principal of "all data" or "no data" by the OSN sites [27]. In this situation, privacy protection becomes a complex issue because user data is migrated into the jurisdiction of another entity. Now, if user leaves or deletes his account from that OSN site, his data will be removed by that particular OSN (at least according to privacy norms, it should be) but what about that third party domains. If user has utilized many applications during his membership of that OSN site, then this data is also available at the servers of these third party domains. Unfortunately, there is no mechanism exists to ensure this down the line private information removal [27]. Other potential source of private information leakage are third party advertisers which are also keep track to the user's activities to perform more targeted or personalized advertisements [27, 28]. OSN sites try to provide anonymous user data to these advertisers. But this anonymous data is also prone to many attacks i.e. re-identification attack. In re-identification attacks, anonymous user data is used to relate or discover with the actual data with a purpose of individual's recognition.

2.6 Preserving user privacy in OSN

Privacy preservation on OSNs is a quite challenging task because of several reasons. First of all, there are a lot of stakeholders with diverse purposes and mottos that cannot be cared with one stone. Secondly, users' private information consists of various formats whereas their familiarity with the trouble and its aftermath is little. Moreover, privacy preferences of OSN users are not same [9, 26]. There is a range of privacy preserving techniques that aim to defend users from few or several privacy intrusions. In this section, we cover a variety of privacy preserving techniques and these counter-measures are divided into technical methods and precautionary measures for better understanding.

2.6.1 Technical Methods

2.6.1.1 Technical solutions for OSN

Although, OSN provides a lot of functionalities to their users but their enormous growth caused many issues i.e. scalability, manageability, controllability and privacy. OSNs are facing attacks mostly from the applications resided on third party application servers and malicious hackers. The most recent technical solution to solve many of these issues is through decentralization. It is also claimed that decentralization by implementing some additional logic will cure not only privacy attacks and security issues of OSNs but it will also improve Internets security overall [42]. Some work is already started in this direction but it did not come out of laboratory to reality [3, 30, 31, 43, 44]. The concept of decentralization is quite straightforward, that is to develop extra layer over the OSN to implement organization oriented privacy preserving logic. According to decentralization advocates, modern OSN sites are mostly suffering from

two dilemmas; information silos and user privacy [43]. Currently, a user cannot share or access his information from one OSN to the other. This issue is referred as information silos. Decentralization could bring following potential benefits;

- Decentralization can improve the privacy and security of the OSN by making it more secure through dedicated privacy mechanism
- Organizations can separate their users by implementing the decentralized version of the OSN. Moreover, they can further use OSN platform to perform their organizational activities.
- Decentralized versions of OSN will be more manageable because of less and same domain users

There are two common approaches to achieve decentralized social network as depicted in Figure 2.9. Firstly, decentralization is performed on single social network service i.e. Facebook which can be termed as single OSN based decentralization. Secondly, decentralization is based on multiple OSNs which can be phrased as multi OSN based decentralization.

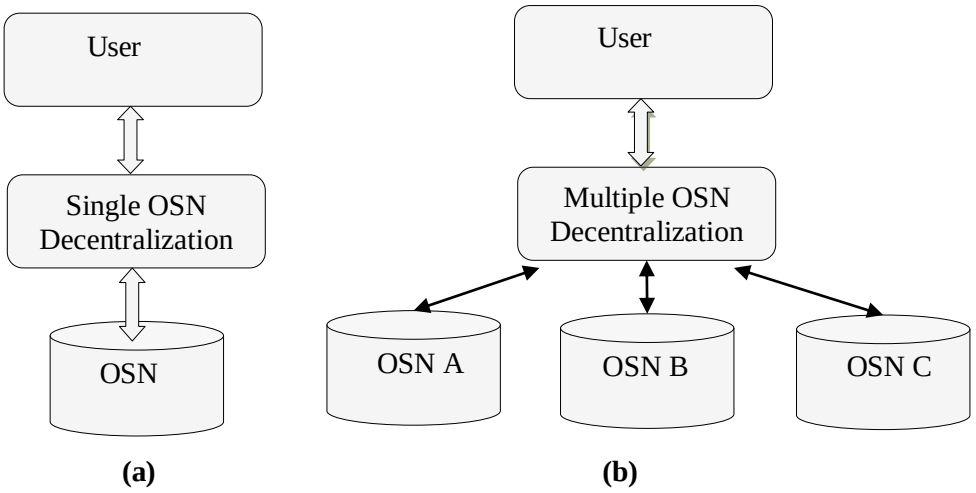

Figure 2.7: Decentralized social networks

Figure 2.7 (a) depicts decentralization where the users of decentralized networks belong to a single OSN site. The decentralized social networks that are discussed in [31, 44] and [3], are instances of single OSN based decentralized networks. On the contrary, if decentralization is based on various OSNs as it is depicted in Figure 2.7(b) then it can be referred as multi OSN based decentralized social networks. Decentralized social network proposed by Yeung, et al. [43], belongs to this type of decentralization. Google took initiative to achieve this kind of so-called distributed online social network (DOSN) where users can share their resources from one OSN to the other without replicating their information. They also developed many Application Programming Interfaces (APIs) to develop decentralized network based on multiple OSN sites with the name of OpenSocial[10]. Whereas, OpenID[11] is a generic id standard that can be used to logon numerous OSNs, simultaneously, for inter-OSN resource exchange. Many OSN sites, i.e. Google Buzz, Yahoo, MySpace, AOL, are following OpenID standard.

Krishnamurthy et al. [27] and Baatarjav et al. [7], give solutions to deal with privacy threats imposed by third party applications. The main idea is to give only required information to the applications and games [27]. Baatarjav et al. suggested a privacy management system for Facebook. They suggested a system that correlates

[10] http://www.opensocial.org/
[11] http://openid.net/

profiles features and network privacy settings. Furthermore, their solution utilizes Bayesian Belief Networks (BBN), a statistical learning method.

2.6.1.2 Technical solutions for OSN users

There are no third party applications that help users to protect their private information on OSN. On the other hand, there exist applications that help users to secure them from privacy threats on Internet and computer networks. These applications include spam filters, antivirus, anti-spywares. In OSN, there exists only vendor provided privacy control mechanism and there is no such application that helps against privacy related attacks.

Some initial research is started in this direction. Studies conducted by Banks et al [4] and Liu et al. [32] suggest techniques to develop such applications. In their study Banks et al, proposed interaction count method to calculate friendship intensity which can further be used for privacy preservation.

2.6.2 Market regulations and Government rules

Market oriented control to preserve user privacy is often sought in capitalistic economies where news or even rumors of customer privacy violation by organizations severely damage company's reputation and even reduce their customer base. In this privacy conscious global society, companies can advertize their strict privacy policy to gain further business [28]. According to Caudill and Murphy, companies have to meet four requirements while collecting user data in European Union(EU)[28];

- The purpose to collect the information should be legitimate and clearly defined
- The purpose should be revealed to owner of the information (user)
- The use of Information should not deviate from original purpose
- The company can keep the data only for the initially defined purpose and if they want to use information for some other purpose then they need to originate new collection and user permission process.

Market regulations are good option if Government does not want to interfere into the market operations. Normally, free competition of markets, self-regularizes these kinds of discrepancies. But, mandatory government rules are necessary in case of monopoly such as Google and iTunes, where end users do not have much options to get some quality service [28].

3 USER PRIVACY CONCERNS

At present, OSNs are major accumulator of user data which give rise to privacy concerns more than ever on the Internet. The main objective of this chapter is to discover users' privacy concerns and compare their worries with their activities on OSNs. This chapter is organized as follows; In Section 3.1, individuals' privacy preserving behavior in real life is compared with their online actions to rationalize the importance of same offline behavior even on OSNs. Section 3.2 provides details about the survey design and its results as well as analysis and conclusions from the results. Finally, Section 3.3 summarizes the responses of the survey for brevity and understanding.

3.1 Introduction

Privacy on Internet is the real concern with the emergence of OSNs, where a variety of user data is enticing different type of malicious users. A number of people feel that there is no privacy on the Internet because it is a public place [45]. Some also claim only small number of Internet users are concerned about their privacy and most do not really care about it. In these circumstances, following questions can be raised: Do individuals really need privacy in OSNs? Do they have privacy in real public places by assuming Internet as a public place? How individuals deal with these privacy issues in real social networks? Do their real life privacy preserving habits conform to their actions on OSNs?

First of all, OSN based digitized social connections carry lot of subtle variations than their offline counterparts. In offline scenario, individuals reveal information according to the context, intensity and trust level of the relationship. The trust or intensity of relationship is a major element for privacy preservation that individuals gain through experience. Furthermore, privacy in online setup is more fragile than the offline because offline environments are not as stabilized as online [25]. Online scenarios further add misery by offering quite robust architecture in terms of persistence, searchability, replicability and scalability than their offline counterparts [25]. Another important aspect is about publically available information, the data that is publically accessible for everyone with no restrictions. People publicize their information on the Internet intentionally or unintentionally. In the first case, they know there information is publically accessible and they did this deliberately to gain potential benefits. In the later case, individuals do not know their information is publically available. That could be because of several reasons such as they are not well familiar to control the access of their information, they are not capable to assess the vulnerabilities or they are not able to use privacy controls. Website's privacy policy is much crucial in the later case. It is their decision, whether to make personal information publicly available by default or it should be restricted (OSN sites are mostly criticized over this issue) and may be changed after user's explicit directions. Important thing is; if something is public then it does not mean people want to publicize it, and "making something that is public more public is a violation of privacy" [25]. Thirdly, this World is now a congenial place to raise instant celebrities because of cheap media for publishing their thoughts, photos, art etc. [25]. This factor allured many people especially teens to post their private data on Internet in order to develop audiences and to gain appreciations.

People may not able to identify potential privacy threats but it does not indicate that they are not bothered about their privacy. Certainly, in some cases individuals expose themselves to gain some potential benefits, but they bear in their minds the costs and benefits of exposing private information. Additionally, individuals may have misperception of being in the real life scenario while acting on OSNs. To further elaborate the issues raised in this section, the rest of chapter covers the details of an

online survey that is conducted to analyze privacy related aspects, concerns, privacy protecting habits and privacy preserving controls provided by OSNs. The discussion about the survey and detailed analysis of the responses are covered in the following sections.

3.2 Privacy concerns survey

3.2.1 Methods

An online survey method is applied to gather Facebook users' views and to analyze their practices with respect to the privacy related issues. Furthermore, we used different tools and techniques to interpret and analyze users' responses. The details of these methods are covered below:

3.2.1.1 Survey Design

A survey website is utilized to conduct the online survey and to investigate users' privacy expectations and concerns. Afterwards, an online questionnaire is produced on that website to execute the survey process. This questionnaire consists of 21 closed questions which offer nearly 300 multiple choice answer options. This questionnaire serves dual purposes of the study where first part inquires about privacy related aspects and second portion of the questionnaire is about users' interaction habits with their reliable friends. The users' privacy concerns are covered in this chapter and analysis of the second part will be covered in the following chapter. The link of this questionnaire is emailed by using different mailing and Facebook interaction facilities. Additionally, we also created a Facebook group and an email account to further help the subjects and to get their views and queries regarding the questionnaire. The questions of survey can be classified as follows:

- Background questions: First three questions are related to the background where information about users' age, nationality and gender is asked. We want to analyze privacy expectations and concerns of individuals that may vary on these factors.
- Internet and Facebook usage questions: The survey contains three questions about the user's expertise levels regarding computers, Internet and Facebook. From these questions, we want to identify users' concerns with respect to their usage and expertise level of Facebook as well as the Internet.
- Privacy concerns as Facebook user: The questionnaire consists of twelve queries related to the user privacy concerns, their privacy preserving habits, and satisfaction over Facebook privacy controls.
- Interaction habits with reliable friends: Final part of questionnaire inquires four questions to identify user's interaction habits with his close and reliable friends.

There are two reasons to use online survey method. First of all, this study is related to a facility that is only available on the Internet. Therefore, users will feel comfortable to fill this survey which does not expect a lot from them in terms of time. Secondly, online survey will expedite the data collection and analysis process. In contrast, the questionnaire is based on closed questions that may reduce users' answers in a certain context but it is required to handle large amount of user views. We try to include every possible answer option in the questionnaire through discussion with Facebook users. To further reduce this limitation, we asked users to send their remarks via email or express their views on a Facebook group created for this purpose. Moreover, we made this survey entirely anonymous to reduce privacy fears of our respondents. The readers are requested to glance at Appendix 7.2 for further details of questions and possible answers.

3.2.1.2 Sample selection

Sample selection for the survey is critical for overall validity and conclusions from results. We decided to conduct this survey with the students and teachers at School of Computer Science in Blekinge Institute of Technology. In Chapter 1, we discussed the reasons and motivations behind this sample selection.

3.2.1.3 Tools, analysis methods and variable selection

We utilized database software, Structured Query Language (SQL) and spreadsheet software to store, retrieve, analyze and visualize survey results. The survey responses are imported into the database and then multiple SQL queries are formed and executed to retrieve data from the gathered survey answers. This retrieved data is then exported to the spreadsheet for visual and graphical representation of the results.

To make the analysis of survey results more concrete, we consider five aspects as independent variables and variations in privacy concerns are analyzed on different values of these variables. These variables include Age (three possible values), Nationality (nearly 200 possible values), Gender (two possible values), Internet Usage **(**four possible values), Facebook usage (four possible values), Friends in the network (six possible values).

3.2.2 Survey results and discussions

3.2.2.1 Demographics

As far as demographics of survey respondents are concerned, 212 individuals participated in the survey and these individuals belong to more than 20 different nationalities that made this survey multinational as well as multicultural. In terms of gender, we received 86% (182) responses from males and just 14% (30) replies from female. This is because of the low female student population in the School of Computing. The participants in this survey are divided into three age groups; teens (younger than 20), middle aged (20-40) and older than 40. Table 3.1 describes the number of male and female participants and their corresponding age group. As it is shown in the table that almost 96% participants belong to the "20-40" age group and we received very little responses from the other age groups. Therefore, where privacy regarding to the age group need to consider, we will mostly reflect on the "20-40" group in our analysis. We will ignore the other age groups, especially "more than 40" class because we have very little participation from this age group.

Table 3.1: Number of Respondents with respect to gender and age

	Younger than 20	20-40	Older than 40
Male	12	168	2
Female	5	24	1

Although we managed to attract near to 25 nationalities in this survey but more than half of the respondents belong to Pakistan. Other major nationalities in the survey are Swedish, Indian and Nepali.

3.2.2.2 Internet and Facebook expertise level

This survey also inquires about the Internet usage and Facebook activity levels of subjects. Most of the users are expert in utilizing the Internet, as it can be noted from Table 3.2, where more than 90% respondents are expert or good Internet users. Additionally, this table also classifies Internet usage levels with respect to the gender and age groups.

Table 3.2: Internet expertise levels with respect to the gender and age groups

	Total	Gender		Age groups		
		Male	Female	< 20	20-40	>40
Expert Internet users	107	95	12	3	103	1
Good Internet users	85	70	15	9	75	1
Average Internet users	19	16	3	4	14	1
Beginner Internet users	1	1	1	1	0	0

Second important factor about the participants, is related to their expertise level of Facebook which they mostly gain by spending time on it. Facebook started its journey from a campus and it is still very famous among students around the World. The purpose of investigating Facebook usage levels is to analyze users' activity and openness relative to their privacy concerns. Table 3.3 describes the respondents' activity levels on Facebook with respect to other criteria such as their gender and age groups. Most of the respondents in this survey are active users of Facebook which can be observed from the table, where almost 80% of individuals are dynamically using Facebook.

Table 3.3: Facebook usage levels with respect to gender and age

	Total	Gender		Age groups		
		Male	Female	Younger than 20	20-40	Older than 40
Very active users	53	42	11	9	43	1
Active users	108	97	11	6	102	0
Rare users	33	30	3	0	33	0
Very rare users	18	13	5	2	14	2

Furthermore, we also considered the Facebook users' activity level with respect to the top three respondents' nationalities which is tabulated in Table 3.4.

Table 3.4: Facebook usage levels with respect to top three respondents' countries

	Swedish	Pakistani	Indian
Very active users	2	35	2
Active users	9	71	12
Rare users	1	24	3
Very rare users	3	10	4

There are two aspects of users' social activity in a real as well as online social setup. Some individuals are active in terms of expanding their network whereas other likes to be active in their small circle of friends through communication, sharing, playing and wishing. In this scenario, number of friends in the respondents' social network is another important factor to judge the former dimension of user activity. Furthermore, this factor is important to compare user privacy concerns and his actions to really preserve it.

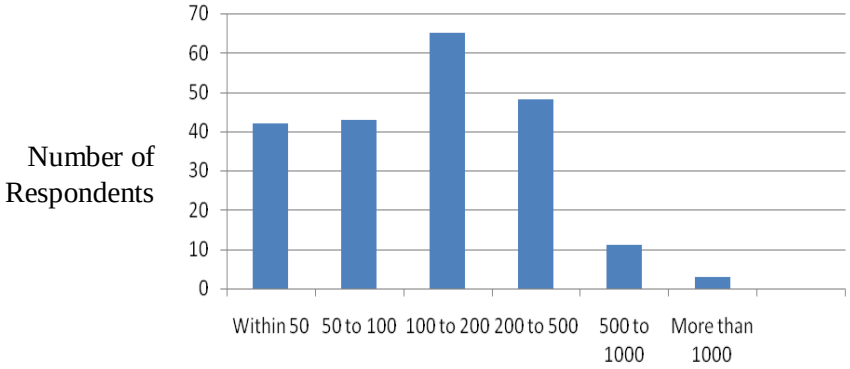

Figure 3.1: Number of friends in the respondents' network

Figure 3.1 illustrates the size of respondents' social network which is grouped into six classes. It is evident from the graph that most users have friends between 100 and 200. This graph also shows that almost 80% of respondents have more than 50 friends in their social network.

3.2.3 Privacy Concerns

We collected varieties of users' views regarding their concerns about their private data on Facebook. The first thing regarding privacy is of course, the number of users who are really concerned to save their private data. As it is claimed by one of industry giant, only 10% of the Internet users are concerned about their private data whereas 90% do not really care about it[12]. This claim is not proved in our study where more than 77% individuals are concerned to save their private data as portrayed in Figure 3.2.

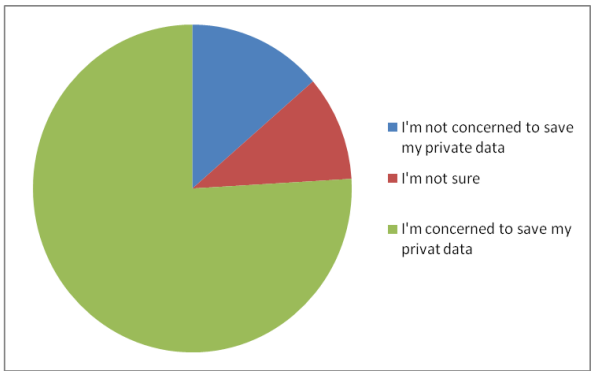

Figure 3.2: Concerns about private data

Moreover, subjects' responses with respect to gender, age and Facebook usage are tabulated in Table 3.5. It is important to note that there is no opinion difference about saving private data among gender level, age groups and usage level of Facebook.

Table 3.5 : Privacy concerns with respect to gender, age and Facebook usage

Concerned about my private data	Gender		Age			Facebook Usage			
	Male	Female	<20	20-40	>40	Very active	Active	Rare	Very rare
Yes	139	22	12	147	2	45	75	26	15
No	25	4	3	25	1	4	19	5	1
Not Sure	18	4	2	20	0	4	14	2	2

[12] http://www.youtube.com/watch?v=pseccQi9ltI retrieved on May, 2010.

OSN Users' Social network size is another interesting aspect to analyze their privacy concerns. Table 3.6 contrasts subjects' responses with respect to their network size where columns show different sized networks. Several interesting facts can be observed from the table. First appealing thing is; the users with very few friends in their network are more concerned about their privacy. This is obvious from the replies of the respondents who include 50 or fewer friends in their network. More than 95% of individuals from "50 friend's group" are concerned to save their private data. Other interesting fact can be seen from the columns of 500-1000 and "more than 1000" groups. These respondents are much concerned to save their private data but did they really doing this practically or they are leaving their private data more at risk. We will examine this type human behavior later in our analysis.

Table 3.6: Replies from different sized friends' networks about saving private data

	Friends in the Network					
	Within 50	**50-100**	**100-200**	**200-500**	**500-1000**	**More than 1000**
Want to save my data	39	30	49	34	7	2
Do not want to save my data	1	8	9	8	2	1
Not sure about my data	2	5	7	6	2	0

3.2.3.1 Information hiding due to privacy concerns

One way of measuring user's confidence on some specific OSN, is by analyzing the number of users who are restraining themselves to expose their private data. In other words, to what extent people are willing the expose themselves on that particular OSN. This information also reflects user's confidence on privacy preserving mechanisms provided by that social service provider. It is strange to observe that 70% of respondents restrain themselves to upload their most private information on Facebook. On the other hand, more than 90% subjects claimed that they do not lie or give false information just to prevent any privacy breach. One reason of this behavior could be that people mostly try to provide correct private information in OSNs because their online networks mostly comprised with their real life contacts.

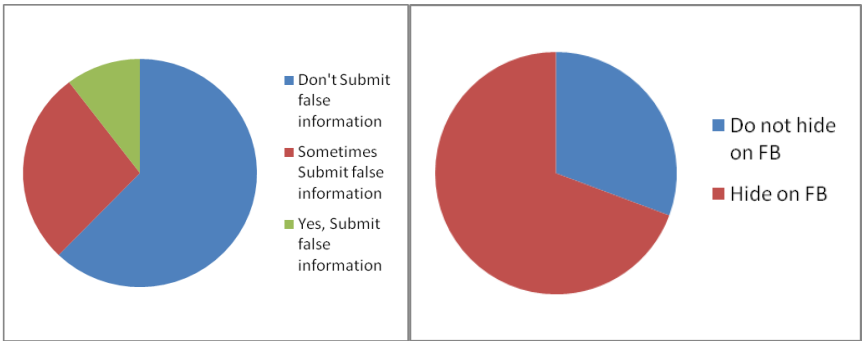

Figure 3.4: Information revelation concerns

Information hiding due to privacy concerns is further analyzed with respect to the gender as well as Facebook activity level and its results are presented in Table 3.7. This table shows the percentage number of respondents who are hiding their information because of privacy fear on Facebook. There are two things that can be observed from the table; firstly, females are hiding more than male. Secondly, people

who are not active Facebook users are more reluctant to expose themselves. In the later case, users' privacy concerns could be one reason for them of using Facebook passively.

Table 3.7: Respondents hiding information with respect to age and Facebook activity

	Gender		Facebook activity levels			
	Male	Female	Very Active	Active	Rare	Very Rare
Hiding Information	68%	80%	74%	62%	82%	78%
Not Hiding Information	32%	20%	26%	38%	18%	22%

We further analyzed, whether the individuals who are hiding their private data can use Facebook privacy settings or not. This analysis gives us an idea about their confidence level on privacy settings provided by Facebook. We explored that 73% of those individuals who avoid revealing their data, are actually using Facebook privacy settings. It is obvious from the last column of Table 3.8. This data shows many individuals do not prefer to reveal a lot in spite of knowing privacy settings.

Table 3.8: Information hiding and using Facebook privacy settings

	Change Facebook setting			
	I don't know whether they exist or not	I know they exist but never used	No	Yes
Hide Information	9	16	14	108
Do not hide Information	5	11	17	32

3.2.4 Privacy preserving habits

According to the Danah [25], privacy related attacks can be averted by following the same real life privacy preserving habits even on the OSNs. Therefore, in order to be more secure one should take care of following things while being online:

- He should be careful while making friends and playing games
- He should effectively use privacy settings to make his/her information as visible or accessible as he/she wants

In this section, we explore that how many subjects are following these simple privacy preserving rules. As far as the care of adding friends is concerned, 66% (34% add if they have common friend and 32% without any factor) subjects add unfamiliar people in their network. There is an element of "trust" in these figures of Table 3.9 where 34% people add unknown individuals in their network only if they have common friends with them. In the next chapter, we will look at how this information can be utilized to calculate friendship intensity. It is also revealed in the table that females are more careful while adding unknown people in their network and only 13% of them add strangers without any trust factor

.

Table 3.9: Adding unknown people

	Age Groups		Gender		Total
	20-40	>20	Male	Female	
Do not add unfamiliar friends	36%	24%	34%	44%	34%
Add unfamiliar friends	30%	47%	34%	13%	32%
Add, only if they have common friends	34%	29%	32%	43%	34%

Using different types of games and applications is an important behavior that describes individuals' approach towards privacy. In this survey, 77% (25% active users) of individuals like to play games or use applications on the Facebook.

The efficient use of Facebook privacy settings also indicates users' privacy preserving habits. As discussed earlier, Facebook only provides information access control facilities for members of one's network and external user, not for games and applications. In our survey reasonable number of Facebook users (67%) replied that they change default privacy setting, this number is much encouraging than the other related work done on Facebook [9, 26, 46]. Most of the users' in our survey belong to computer science background and it could be one reason for that number to be so high.

Table 3.10: Changing default privacy setting on Facebook

Changing Default Privacy Settings	Number of people
People who don't know there exist such settings	14
People who know they exist but never used	27
People who never changed	31
People who change privacy settings	140

We also found in our survey that changing Facebook default privacy settings also depends on the Internet expertise level. This aspect is much apparent when we further dig into the data as shown in Table 3.11. This table shows that 75% of expert Internet users change default privacy settings.

Table 3.11: Changing default privacy settings (Internet Expertise levels)

Expertness Level	Change Settings	Don't Change
Average	58%	42%
Expert	75%	25%

We also compared Facebook usage levels with altering the privacy settings. It is revealed in Table 3.12 where active Facebook users are actually utilizing its privacy control settings more than the less active users.

Table 3.12: Changing default privacy settings (Facebook expertise levels)

Expertness Level	Change Settings	Don't Change
Very Active	75%	25%
Active	73%	27%
Rare	45%	55%
Very Rare	33%	67%

3.2.5 Privacy threats

OSN users face multiple privacy threats which range from external corporate level intrusion to the violations by their close friends in their own social network. This survey is mostly related to the internal privacy threats.

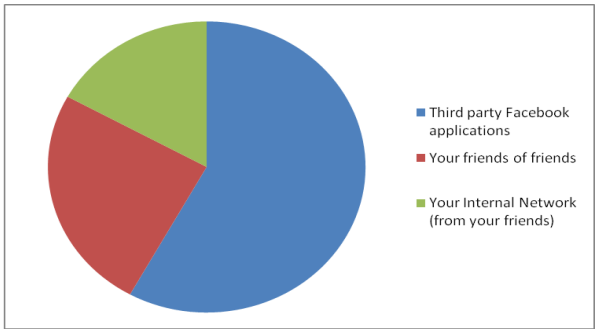

Figure 3.3: Privacy threats

Figure 3.3 illustrates subjects' reponse regarding potential internal privacy threats. This figure illustrates that more than 50% users feel third party applications and games as their biggest privacy threats. It is interesting to note that 77% of individuals who said Facebook third party application are biggest privacy threats for them, are also using them.

The friends of one's own network could also be a serious privacy threat. Especially, the careless habit of including strangers in the social network can put not only the data of that particular person at stake but various information of all others in his network. Furthermore, malicious members of one's own social network are major cause of identity theft and social engineering[13] attacks. In the survey, 66% people think that some of their friends could be a privacy threat for them. Among these 66%, 28% categorically said that there are some malicious users in their social network.

Table 3.13: Internal privacy threat and privacy preserving habits

	Add strangers	**Add stranger if they have common friends**
Friends can be privacy threat	28%	41%
Friends may be a privacy threat	35%	34%

Table 3.13 provides another example where people know that one careless act could risk their private data but they are just doing it, strangely. As it can be observed from the table, 28% of the subjects among those who strongly believe that some of their friends could be a privacy threat, are just keep on adding strangers in their social network. Another factor to estimate user's confidence on their social network is through analyzing their willingness to expose their private data to the members of their own network. As far as exposing their private data is concerned, 87% of people only want to expose their private data to some of the friends in their nework. The willingness of information revelation at different levels is depicted in Figure 3.4. This information does not signify that these individuals are actully following this habit (showing their data to only selected friends) on Facebook but it is more like they want to.

[13] http://en.wikipedia.org/wiki/Social_engineering_(security)

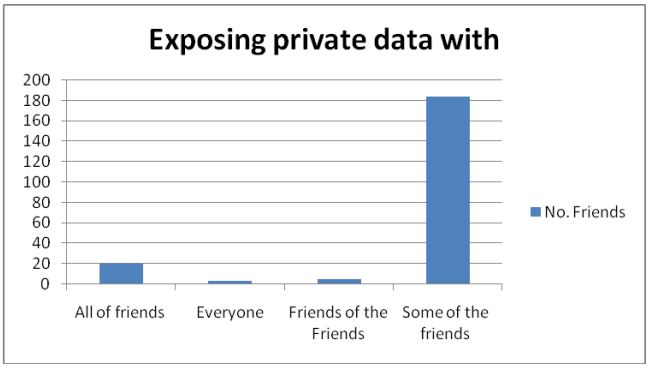

Figure 3.4: Different levels where respondents want to reveal their private data

3.2.6 Concerns over governmental interference

According to the Facebook privacy policy Government agencies can access or Facebook can provide, the data of Facebook users due to security issues. In this survey, 43% of the subjects strongly condemned this interference by the government. On the other hand, 20% of the respondents support this act by the higher authorities whereas 39% people feel that this act has no harm for them.

3.2.7 Facebook privacy settings

Finally in our survey almost 56% respondents feel that the Facebook privacy settings are difficult to use. This is quite thought provoking response in the circumstances where most of the subjects have computer background. If they feel, the privacy settings are difficult to use then it will perhaps even more difficult for the other users. We already mentioned that 67% of individuals are using Facebook privacy settings.

3.3 Summary of survey results

Privacy protection is an endeavor which should be done by the both, the data owner (member of OSN) and the data collector (OSN), hand to hand. The user should be educated and made realized about the gravity of the situation, especially by the OSN service providers. In this study, we conducted a survey to analyze users' understanding about privacy related aspects and concerns regarding their personal data. This online survey is consisted of more than 20 closed questions related to the Facebook. We managed to get little more than 200 responses from the individuals who belong to around 25 different nationalities. The age of the most survey respondents is between 20 to 40 years and almost 90% of them are actively using Facebook and the Internet. Near to 60% respondents have more than 100 friends in their network.

In this survey, we strived to identify users' privacy concerns through a variety of inquiries regarding user privacy, information hiding, and information lying. Almost 80% of subjects, with same proportions with respect to age, gender and Facebook usage level, want to preserve their private data. Furthermore, individuals with fewer friends in their network are even more concerned about their private data. Information hiding is another factor that shows privacy concerns of users and 70% of individuals hide their private information due to privacy leakage fears on Facebook. We also observed that female respondents and "less active" Facebook users are hiding more than their corresponding groups. Moreover, 73% of those who are hiding their private information are also using Facebook privacy settings. In this survey, 90% of individuals respond that they do not lie to protect their private data.

OSN users' activities indicate their familiarity as well as seriousness regarding preservation of their private data. We observed 66% individuals add strangers in their network but female respondents are relatively more careful while adding unknown people in their network. More than 70% individuals like to play games and applications on the Facebook and 70% among them also feel this facility as major

privacy threat. We also noticed that 67% of individuals changed default privacy settings which indicate that respondents are much familiar with the privacy preserving mechanism provided by the Facebook. In order to find out, whether the use of privacy settings depends on users' activity level or not, we found active Facebook and Internet users are relatively more familiar with the privacy settings.

Most respondents (58%) feel that Facebook third party applications are the biggest threat to their privacy. Rest of the subjects think that their friends (16%) and friends of their friends (26%) can violate their privacy. While highlighting the internal privacy threats we found, 66% of individuals feel that their could be some malicious friends in their network. Its also interesting to note that, 28% of those who said friends could be a privacy threat, also like to add unknown people in their network. Individuals willingness to expose their private data in front of their network is another factor that reflects their confidence level on their own network. Near to 90% individuals want to share their private data only with selected friends in their network. In addition to internal privacy threats, 43% of individual also condemned privacy interference by the Government agencies. Finally, as far as the usability of Facebook privacy settings is concerned, 56% of individuals feel that Facebook privacy settings are difficult to use.

4 FRIENDSHIP INTENSITY CALCULATION

Unlike real life friendship, most OSNs only allow to make a binary type of friendship where an individual either is a friend of another individual or not. This kind of adamant friendship definition brings numerous issues in OSNs, mostly related to the privacy. In this chapter, we argue to incorporate a functionality with current OSNs that automatically calculates the relationship strength between individuals by using their interaction data, and other metrics such as mutual friends and profile visits. The rest of the chapter is organized as follows: Section 4.1 emphasizes the importance of having a friendship intensity calculation feature in OSNs. A variety of metrics can be used to calculate friendship intensity in an OSN and Section 4.2 explores these factors along with their related complexities. One of the major contributions of this work is the use of data mining to calculate friendship intensity/levels, which is covered in Section 4.3. The details of the data mining experiment are discussed in Section 4.4. Finally, Section 4.5 proposes a framework that utilizes predicted friendship levels to improve the user privacy in OSNs.

4.1 Introduction

Throughout the text, we discussed that individuals are connected with relations and these ties form the basis of their social network. Furthermore, the social network does not solely depend on an individual node but also on the connections possessed by that particular node. These connections can be characterized by content, direction and strength [16]. The intensity of a connection is also termed as the strength of that relationship. This characteristic of a relationship indicates the closeness of two individuals or how powerfully two nodes are connected with each other in their social graph. Moreover, the strength of the relationship is a mental state which can also be perceived as levels of a relationship such as good, average or bad. In real life social networks, the friendship strength is a crucial factor for individuals while deciding the boundaries of their privacy. Moreover, this subjective feeling is quite efficiently utilized by human to decide various other privacy related aspects such as what to reveal and who to reveal.

On the other hand, online connections can also be classified into “strong” and “weak” ties, but relationship intensity and its context is not symbolized in most OSNs. Moreover, individuals follow different approaches in order to make online friendship. A few people indicate anyone as a friend, while some stick to more conservative definition of friendship, and most list anyone as friend who they know or not totally dislike [26]. In these situations, most OSNs evolve a different type of friendship phenomenon, where one may not trust or even be acquainted with his “friend”. It is also observed, OSNs are mostly helpful in preserving so-called weak ties and, no major accrual in strong connections is detected in online scenarios, instead [25]. The underlined scenario make online network very much like an “imagined community” [26]. Therefore, in addition to other privacy and security threats, individuals can also face privacy threats from their own social network members due to the lack of trust and acquaintance. OSN users are unable to control these privacy vulnerabilities because:

- Not enough privacy control facilities are provided by OSNs
- The users do not know they have these facilities
- The privacy controls are difficult to use
- Friendship is only type of relationship provided by most OSNs to establish a connection between individuals
- Individuals are unable to identify potential privacy leakage connections

Above all, one major issue with the OSN architecture is that they do not consider the intensity (how good friend) and context (class-mate, teacher, colleague, subordinate) of a relationships. Unusually, an OSN user has to send "friendship" requests even to their parents, relatives or mentors for including them into their social network. Recently, some social networks start providing facilities to control information access but they are difficult to maneuver and normally overlooked by the users. Furthermore, the relationship status between individuals tends to grow or deteriorate with the passage of time. Therefore, these privacy settings once set, may become meaningless after sometime. The binary nature of relationship makes privacy much uncontrollable for OSN users. In these circumstances, the estimation of friendship intensity is quite useful to identify internal privacy threats.

Friendship intensity information can further be utilized to improve other ONS based applications. The other benefits include; the calculation of trust level between individuals, improvements in the recommendation process (e.g. friend recommendation on Facebook) and improvement in SNA. This part of thesis introduces the metrics to calculate the level of a relationship as well as purposes a novel approach for identifying friendship levels by using data mining techniques.

4.2 Factors to calculate friendship intensity

OSNs have several indicators that can be used to predict the friendship intensity between individuals. In general, these factors can be divided into interaction based and non-interaction based metrics. These factors can be used separately or in combination with each other.

4.2.1 OSN Interactions

OSN sites provide a variety of interaction, sharing and communication facilities that include real time as well as non-real time interactions as illustrated in Figure 4.1. The real time interaction styles require the presence of interacting parties e.g. chatting, video conferencing, game playing etc. In addition to the real time interactions, OSNs also offer a variety of non-real time communication styles e.g. private messaging, blogging, comments, compliment, status updates. OSN private messaging is comparable to traditional email but unlike traditional email, it reduces the factor of "any to any communication" by granting more control over incoming messages which normal email do not grant. Furthermore, user can also write messages on the profile of their friends in the form of comments. This type of interactions can be referred as public messages because they are visible to anyone who visits the profile of that particular person.

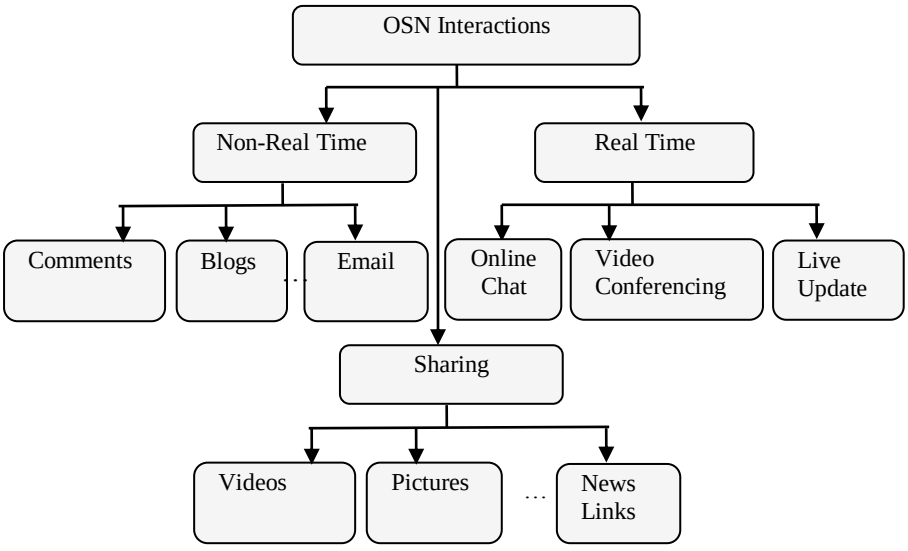

Figure 4.1: OSNs interactions classification

These interactions are main source to estimate friendship intensity in OSN context. There are several interaction aspects that can be considered to identify friendship levels or strength. These factors include type of interaction, actual information and context of interaction. Furthermore, these features can be applied in various ways for estimating friendship strength. A number of interaction based techniques are covered below:

4.2.1.1 Interaction type

The type of interaction is quite important in order to calculate friendship strength because numerous interaction facilities are provided by OSNs. Individuals choose an interaction type according to the nature of information resource, type of relationship and its target audience. For instance, private messaging is a preferred communication style if information is a secret or only concern to one person in the network. On the other hand, wall posts are normally considered, if information in the interaction is concerned to the whole network. Therefore, interaction type defines the intimacy, openness, sensitivity as well as the strength of a relationship between communicating parties.

The survey respondents have selected private messages, comments and chatting as the most preferred interaction styles in order to communicate with their most reliable friends. It is showed in Figure 4.2, where private messaging is selected as the most common interaction style for communicating with reliable friends. Moreover, 68% of the subjects who has selected private messages, also utilize "comments" as second preferred interaction type whereas, 66% of the subjects have selected chatting as the third preferred communication style.

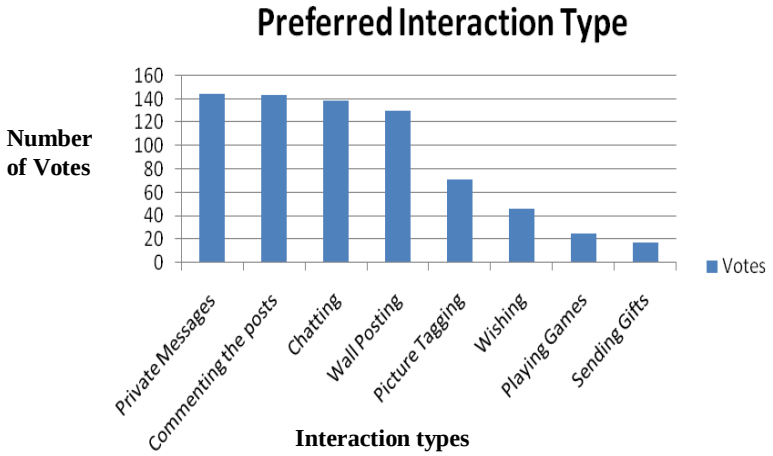

Figure 4.2: Preferred interaction types

There are two motives of considering interactions selectively; First of all, we cannot include all interaction types for calculating the friendship intensity due to efficiency and privacy concerns. Furthermore, the importance of interaction types varies from user to user as shown in Figure 4.2. To manage the later issue, numerical weights can be attached with an interaction to increase/decrease its contribution in friendship intensity calculation.

4.2.1.2 The interaction count

A simple count of interactions is one way of avoiding the complexity and diversity of an interaction. This method of calculating friendship strength, is suggested by Banks et al. [4] and Musial [15]. Interaction count refers to the total number of

interactions between a user and his friends within a certain period of time. In Equation 1, $T_{interaction}$ represents the sum of all interactions of type *I* between a user *u* and his friend *v* in a certain time *t*. Interaction count method argues to count interactions of all types whereas the method described in Equation 1, also prefixes weight with each interaction in order to increase or decrease its contribution in prediction. The empirical justification of increasing/decreasing the importance of a specific interaction is provided in the preceding section.

$$T_{interaction}(u,v) = w_i \sum^{t} I_i(u,v) + w_j \sum^{t} I_j(u,v) \ldots + w_n \sum^{t} I_n(u,v) \qquad \textbf{(1)}$$

Several factors can affect interaction count while calculating friendship intensity through this metric. First of all, interaction habits of individuals are not same with their reliable friends. A number of individuals like to interact with their strong friends frequently, and rarely interact with their weak ties or vice versa. Secondly, the context of an individual is another important influencing factor for interaction count metric. Individuals do not tend to interact a lot if they have same context such as they are working or living in a same office, house, and city. Finally, the user activity levels also influence individuals' interaction behaviour with their friends. Normally, less active OSN members do not frequently interact with their friends and in response they are rarely contacted by others.

Apart from these issues, 70% of the respondents prefer to interact with their reliable friends on Facebook, in our survey. Table 4.1 illustrates number of respondents who like to interact with their reliable friends on Facebook.

Table 4.1: Interaction likeness with good friends

Like to interact with reliable friends	Respondents
No	3
Only when it is necessary	60
Yes	147

To analyze whether user's interaction habit depends on his activity level, we further dig into the survey data and found supporting results which are illustrated in Table 4.2. According to the table, 81% of very active users prefer to communicate with their good friends on Facebook and this ratio is higher than overall ratio of 70%. Moreover, 77% of the active Facebook users prefer to communicate with their good friends and this number is little less than the very active users. Finally, for rare Facebook users this ratio is reduced to 44%. This survey data supports our argument that user interactions with his good friends depends on his activity level in that particular OSN.

Table 4.2: Interaction habits with respect to Facebook activity level

	Facebook usage levels		
	Very active	Active	Rare
Do not like to interact with good friends	1	2	0
Only interact when it is necessary	9	23	28
Like to interact with good friends	43	82	22

The other important factor is the estimation of total interactions, which an individual can perform with his friends in a certain period of time. This factor becomes even more important if we want to develop artificial training data to develop a data mining model. Table 4.3 shows an indicator for that estimation which is inquired in the survey. The table data reveals an important feature regarding interaction count, where 85% of individuals interact with their good friends at least once or many times in a week. In the next section, we discuss how this estimation could be helpful in setting minimum bound for interaction based attributes in the process to create training data.

Table 4.3: Interaction count with good friends

Interaction count	Responses
Many times in a week	134
Once in a week	47
Many times in month	14
Many times in a year	8
Once in month	8
Once in a year	1

4.2.1.3 The content and context of interaction

Friendship intensity can be calculated more accurately by understanding the contents of the interaction through Text Mining [47] and Natural Language Processing (NLP) techniques [48]. This is more powerful method which reveals actual nature and purpose of the interaction. We can apply a text classification technique to build a classifier which takes all interaction of a specific type and categorize them into good interaction, bad interaction or fair interaction. Finally, relationship intensity can be calculated by simply counting and taking ratios of these counted interactions. This type of interaction content based classifier is portrayed in Figure 4.3, where classifier takes all interactions of type *i* and classifies them into good, fair or bad.

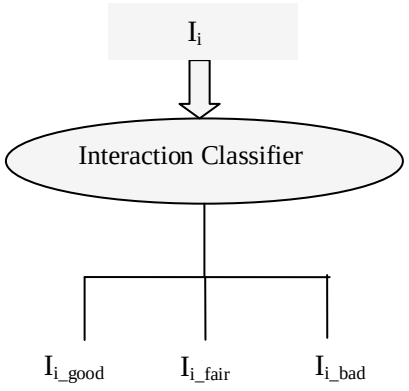

Figure 4.3: Interaction classifier

Furthermore, we cannot simply divide an interaction into these levels unless the context of discussion is obvious. For example, while communication, if someone writes “very bad” in an interaction. Subsequently, we cannot categorize it as a bad interaction by just considering it as a remark on communicating individual rather it can be a comment on some issue. Interaction context is related to the actual subject of interaction which may comprise several previous interactions. The whole thread of

interactions has to consider for understanding the context of the discussion. This is a complex way of classifying friendship intensity and is out of the scope for this thesis.

4.2.2 Mutual friends

Apart from the interaction based indicators defined in the preceding section, common friends between individuals, is a basic trust factor that can be utilized in friendship intensity calculation as well. Mutual friends refer the total number of common friends between two individuals in a social network. Many common friends lead to the fact that individuals are strongly connected with each other, or they may have same context. This information can be used as friendship intensity calculation metric separately or with some other criteria. The importance of mutual friends as a factor is further highlighted in our survey where 46% of individuals add strangers in their network only if they have mutual friends.

$$T_{common}(u,v) = T_{friends}(u) \cap T_{friends}(v) \quad \textbf{(2)}$$

In Equation 2, T_{common} denotes common friends between *u* and *v* which can be calculated by taking intersection of *u's* friends and *v's* friends. This information can be used in several ways, for instance, we can simply calculate the contribution ratio of this metric in decision making. Equation 3 describes a simple way of integrating mutual friends with some other criteria such as interaction count. In this equation, the value of T_{common} could be one or zero depending on the existence or non-existence of mutual friends between *u* and *v*.

$$T_{common}(u,v) = \begin{cases} \mathbf{1}, & u \text{ and } v \text{ have atleast one common friend} \\ 0, & \text{otherwise} \end{cases} \quad \textbf{(3)}$$

4.2.3 Profile visits

This study has identified "profile visits" as another indicator for friendship intensity calculation. Profile visits refer to the total number of times an individual visits the profile of a certain friend for some purpose e.g. to check his activities. This metric is further justified through our survey as depicted in Table 4.4 where almost 80% of subjects frequently like to visit the profile of their good friends.

Table 4.4: Profile visits

Profile Visits	Respondents
No	44
Yes	168

In Equation 4, profile visits function counts total number of times a certain individual *u* visits the profile of his friend *v* within a certain period of time.

$$T_{visits}(u,v) = \sum^{t} P(u,v) \quad \textbf{(4)}$$

4.3 Friendship intensity using data mining

Data mining is a sub area of Machine Learning which instructs a model by using training data and utilizes this trained model to solve real life problems [11, 49]. Data mining is a holistic process, not a single algorithm or technique which consists of several phases. In general, the process of friendship intensity/level calculation through data mining is portrayed in Figure 4.4 which comprises offline and online modes. In offline mode, training features' selection, training data preparation, pattern discovery

and evaluation phases are performed. This process is referred as data mining model creation or training phase. This model is then incorporated with some particular OSN to calculate friendship intensity in online mode.

In offline mode, the first phase is the selection of appropriate features or attributes for the training of data mining model. There is a variety of features that range from interaction based features to profiles visits for calculating friendship intensity. Most of these features are based on interactions such as interaction type, interaction contents and context. In addition, other features such as current status, user activity level, profile visits can also be considered separately or they can be integrated with interaction based features. Attribute selection is an important phase that should be performed carefully.

Data mining process enters into the next phase of training data preparation after selecting different attributes. In this phase, example instances of selected attributes are collected naturally or artificially. In the former case, actual historical data is used for the training of data model. If actual data for training is not available then it is made artificially using real assumptions. This kind of training data can be referred as the artificial training data. Every method of training data creation has its own pros and cons [11]. Training data preparation is perhaps the most critical phase in data mining process because the performance and accuracy of data mining model solely depends on its training data.

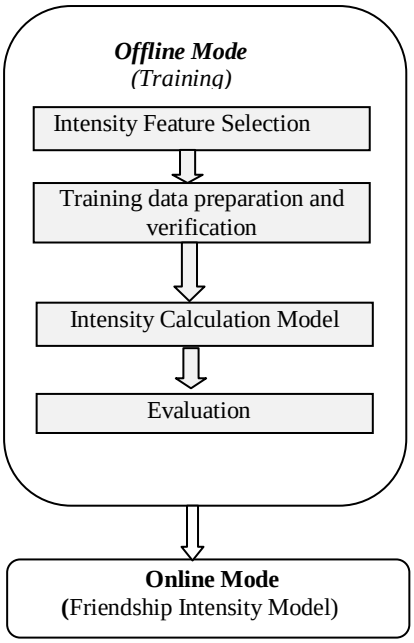

Figure 4.4: Friendship intensity calculation using data mining

The process of attribute selection and training data preparation is further illustrated using the hypothetical training data of Table 4.3. The rows of the table represent 14 training instances showing some user's (suppose X) anonymous friends. First of all, six features or attributes are selected that are illustrated in the first six columns of the table. These attributes include five interaction based attributes such as private messages, chats, wall posts, comments and one current status attribute. The current status is a binary attribute which shows; in case 1, the X and his corresponding friend have same context which means, they are living or working in same home or organization and 0 otherwise. The numerical values of interaction based attributes show the total number of interactions of specific type in certain period of time. The last column represents percentage value that, the model should predict in case of different combinations of other attributes. For example, in the third row an instance is

provided to the model where it should predict 75% of friendship intensity with corresponding values (18, 36, 18, 12, 10, and 0) of other attributes. We can also state, X has 75% friendship strength, if some of his friend has this number of interactions and status value. This type of supervised data mining is referred as numerical prediction [11].

Table 4.5: Hypothetical training data

Messages	Comments	Chatting	Wall Posts	Tagging	Context	Intensity
30	60	30	15	10	0	0.99
22	46	23	11	8	0	0.87
18	36	18	12	10	0	0.75
15	30	15	9	6	1	0.73
14	29	14	8	5	0	0.61
13	28	13	7	4	0	0.57
8	23	8	7	3	0	0.41
9	22	10	8	4	0	0.38
10	21	11	7	5	1	0.44
8	20	8	4	3	1	0.4
4	12	4	2	2	0	0.21
2	4	3	2	2	0	0.1
3	2	4	3	2	1	0.15
3	2	4	3	2	0	0.06

In addition to the numerical prediction we can also make the training set of Table 4.3, a data mining classification problem [11] by replacing the numerical values of last column with categorical classes such as very good, good, average. In that case, we can label this column as the friendship levels attribute, and it will predict the friendship levels rather than numerical values. For instance, we can take five friendship levels such as very good, good, average, low and very low.

In the offline mode, the next phase is data mining model creation, where a model is generated from the training data using various data mining algorithms [11, 49]. The use of these algorithms depends on the nature of the problem and characteristics of the data in the training set. For the friendship intensity problem, we can use classification as well as numerical prediction algorithms [49]. Data mining model creation process is further illustrated through the following regression Equation 5. This equation is calculated from the training data of Table 4.3 by using Weka [11].

$$\begin{aligned} \text{Intensity} = {} & 0.0436 * \text{messages} + 0.0093 * \text{comments} + \\ & (-0.0361 * \text{chatting}) + 0.0044 * \text{wall\_posting} + \\ & 0.0211 * \text{tagging} + 0.0726 * \text{context} + 0.0374 \end{aligned} \quad \textbf{(5)}$$

The above data mining model is a simple regression model that predicts different friendship intensity values between the ranges defined in the training data. In this data mining technique, an equation is generated which fits the training dataset. There are independent attributes which construct a resultant equation after taking together. We gave the outcome values as intensity from 0.99 to 0.01 depending on the independent attributes.

Besides generating different strength values, we can also use classification algorithms to generate different levels of friendship as illustrated in Figure 4.5. In this figure a decision tree classification model is generated.

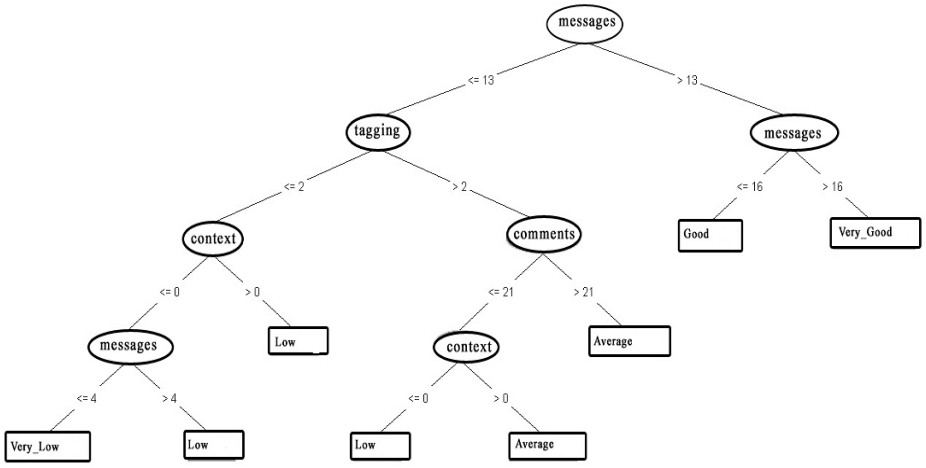

Figure 4.5: Decision tree for friendship classification

In Figure 4.5, a decision tree is constructed by modifying the last column of training data in Table 4.5 with levels rather than numeric values. The numerical values are changed as: Very Good (0.99 ~ 0.85), Good (0.84 ~ 0.70), Average (0.69 ~ 0.40), Low (0.39 ~ 0.25), Very Low (0.24 ~ 0.01). This decision tree can be utilized to identify the levels of friendship on the basis of different values of interaction attributes.

Once the data mining model is created, the next step is its evaluation. This model is evaluated using different statistical techniques [11, 49]. Finally, this data mining model can be integrated with the applications after the evaluation of the model.

4.3.1 Pros and cons of using data mining

In this problem domain, data mining techniques could bring several benefits such as learning and efficiency. In addition to this, we have a couple of motivations in selecting data mining. First of all, data mining techniques provide a natural way to integrate several features/metrics in order to calculate friendship intensity. We simply provide different instances of these features and data mining model learns according to these training instances. Secondly, data mining approach is quite adaptable in this scenario. It signifies that we can train one general data mining model and this model can be integrated with every OSN user, initially. Later on, this model learns according to the preferences of some specific user. In addition to these benefits, the accuracy of data mining model largely depends on the training data. The data mining models based on inadequately designed training data may face some serious accuracy issues.

4.4 The experimental procedure

The objective of this experiment is to determine the feasibility of applying data mining techniques to predict the friendship levels. Furthermore, this activity also evaluates the performance of two classification algorithms in the process. The experiment is conducted using Weka [11] workbench, an open source data mining tool.

4.4.1 Data set

A good training dataset must consider maximum features of the interactions as well as non-interaction based features. Moreover, the training data set should also cover the maximum and minimum values of selected interactions. In the survey,

questions regarding the communication behavior of users with their good friends are asked to estimate maximum and minimum interaction values. The 64% of the respondents said that they interact with their close friends many times in a week whereas 22% of the people said they interact with their close friends once in a week. From that information, it can be inferred that a total of 86% of subjects interacts with their friends at least once in a week. We can assume minimum and maximum values through these results. The second task is to define the time span in which a good count of interactions can be supposed, since the relationships are very dynamic in nature. They can become stronger and weaker with the passage of time. We have selected five months as a reasonable time period. In this time period, the weakening or improvement in a relationship is easily observable. The minimum numbers of interactions between close friends are assumed twenty in five months by taking at least one interaction in a week. Other important aspect is selection of attributes or features which is discussed in the 1st chapter.

The next task is to generate data in normalized form. The respondents gave their votes to frequently used interaction types. For the purpose of normalized generation of training data, it is necessary to select the maximum values of interactions according to the user's response. The maximum and minimum values of each interaction for the training data set are chosen through survey results. The total interactions of all types are one hundred and forty in five months by assuming maximum one interaction per day.

The training data is randomized in a manner where the sum of all interactions of any type cannot be greater than 140. The maximum value for each interaction is calculated by multiplying the minimum value of that particular interaction. This process is needed to maintain the lowest level of close friendship, with the voting ratio of that interaction and then dividing with the sum of all voting ratios of selected interactions. In Equation 6, I[t] refers to some particular interaction.

$$\text{Count}(I[t]) = \text{MaxCount}(I[t]) * \text{VotingRatio}(I[t]) / \sum \text{VotingRatios}(I[t]) \quad \textbf{(6)}$$

Table 4.6: Selected attributes and their votes

Types of Interactions	User's Votes	Voting Ratio
Messaging	150/210	0.71
Commenting	135/210	0.64
Chatting	146/210	0.69
Wall_Posting	141/210	0.67
Tagging	74/210	0.35

The sum of all interactions should not exceed 140. In this way, we have reduced the size of our training data. A non-interaction attribute, profile visits, is added in the training data set. The motivation is that 80% of the total number of respondents said that they like to visit their close friend's profile. A fix value is given to the visiting profile attribute which award its weight equal to 4 interactions. The next task is to define the levels to which each instance belongs. It has been observed that there are two types of relationships; one is the acquaintance which means that user knows some person but he has no direct concern with him and second relationship is the close friends. Friendship levels are divided into five levels that include very good, good, average, low, and very low. These levels are assigned to the instances according to our survey responses where at least 20 interactions are necessary to qualify for the average friendship level. The intensity is calculated for every instance according to the following Equation 7. It is the sum of all interactions plus the user's likeness to visit his friend's profile.

(7)

$$Value = \sum I[t] + 4 * (0, 1)$$

Finally, the class values are labeled using the ranges defined in Table 4.7. This table describes various friendship levels according to the accumulated values of all interactions.

Table 4.7: Assignment of levels in training data

Level	Value
Very_good	> 70
Good	> 45 and < = 70
Average	> 20 and < = 45
Low	>10 and < = 20
Very_low	< = 10

4.4.2 Algorithms

J48 and Naive Bayes are selected in order to investigate the usefulness of the learners. The motivation of selecting these algorithms for experiment is that these algorithms are fundamental and belong to the different families. Moreover, the purpose of this experiment is not to compare these algorithms but to check whether the training data is classifiable or not. Therefore, we have selected relatively less complex algorithms for this classification task. This experiment process will also demonstrate the use of classic data mining algorithms for friendship level prediction. This experiment is performed using the default configuration of these algorithms.

J48 is a decision tree based learning model that predicts the target value of the new instance based on the attribute values available in the training data set [11]. In decision tree, the internal nodes represent the attributes and the branches represent the possible values of the particular node observed in the training data. The end nodes predict class value to which the instance is belonged. The predicted class attribute is called the dependent attribute because its value is dependent on the value of other available attributes. The decision tree is build by identifying the attribute that distinguishes the various instances clearly [11]. This attribute has the highest information gain because it tells most about data that classify the best. If any of the possible values of this attribute classifies all the instances into the same target value, then this branch is terminated and assigned the obtained target value [11]. For the rest of the values, we find another attribute with the highest information gain. This algorithm continues until it gets a combination of attributes that gives a clear decision for a particular target value [11]. From this constructed tree, we can predict the classes of new data instances by following the attributes.

The other algorithms, Naïve Bayes works on a simple, but rather spontaneous concept. In this simple technique, all the attributes are considered to be of equal importance in decision making and are independent to each other [11]. Although, it is unrealistic approach but this technique performs well in comparison to various other complex algorithms. This algorithm follows the rule of conditional probability [11].

4.4.3 Evaluation

The performance of these classifiers are evaluated using the 10 fold cross validation (CV) [11]. Cross validation is used in the situations where limited number of data instances are available which have to use for both training and testing of the model [11]. CV ensures that the model is tested only by using those instances which are not part of the training. For that purpose, all the data is partitioned into n folds where only one fold is used to test the data and the rest (n-1) of the folds are utilized in the training process. This process is iterated n times and average error of the model is estimated. We partitioned our training data into 10 folds. The extensive use of 10 folds

on different data sets have shown that 10 is the best number of folds to attain the good estimate of accuracy [11].

We have selected the accuracy metric (the number of correct classifications divided by the total number of classifications), root mean squared error and mean absolute error for the evaluation of classifiers. The basic formulas to calculate these metrics are given below:

root mean-squared error $$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}} \quad \textbf{(8)}$$

mean absolute error $$\frac{|p_1-a_1|+\ldots+|p_n-a_n|}{n} \quad \textbf{(9)}$$

The predicted values are *p*1, *p*2… *p*n and the actual values are *a*1, *a2*… *a*n. *P*1 is the probability that a particular prediction is in *i*th class. The classifier performs well, if its root means squared error value is near to 0 [11]. The mean absolute error is an alternative evaluation measure which is an average of individual errors without considering their sign [11]. The Mean-squared error tends to exaggerate the effects of outlier instances whose prediction error is larger than the others but absolute error does not have this effect because all error are treated equally according to their magnitude [11]. A classifier performs well, if there is not much difference between its mean absolute error and root mean squared error.

Table 4.8: Comparison of classifiers

Measures	J48	Naïve Bayes
Correctly Classified Instances	292	324
Incorrectly Classified Instances	112	80
Mean absolute error	0.1176	0.1293
Root mean squared error	0.3121	0.252
Total number of instances		404

The Naïve Bayes algorithm performs well on training set and its success rate is 80.19%, whereas the success rate of the J48 is 72.27%.

4.5 Framework for utilizing friendship levels

According to the survey response, 80% of the people want an automatic or semi automatic privacy preserving mechanism. In addition to this, there are following reasons of having this type of mechanism:

- Most of the people hide their personal data because of the lack of trust on the current privacy control mechanism
- Some of the people are in a habit to add strangers in their network which can be privacy threats to them.
- People with the average or low Internet usage find the privacy settings difficult to understand or have less knowledge to use them.

A comprehensive framework which leverages the benefit of identified friendship levels is proposed in this study. Figure 4.6 figure describes this semi-automatic privacy control framework.

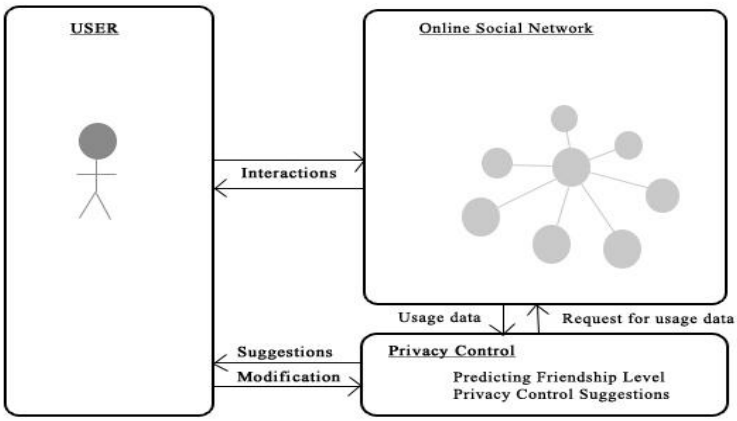

Figure 4.6: Privacy Control Framework

User interacts with their friends using different means. All the communication among one's network is observed and stored in OSNs. The privacy control layer sends request to the OSN site for the particular user's usage data. This layer further defines the friendship levels through the usage data. This layer will suggest the different privacy settings according to the predicted friendship levels. Users can modify friendship levels as well as suggested privacy settings. The following steps describe the working of proposed framework briefly:

Step1: Users interactions are counted after a predefined time period.

Step2: These interactions are further given to the learned data mining model which will classify the friendship relations into predefined levels

Step3: After classifying the friends into levels, the system will suggest different privacy settings for each level

Step4: The user is asked to manipulate the groups if he thinks that any of his friends is incorrectly classified. He will be able to change his friend's level. This manipulation will be saved with the corrected classification and will be used in enhancing the performance of learning model.

Step5: The privacy settings will be changed according to the user's response.

5 CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

This research has investigated several privacy issues as well as methods to calculate friendship intensity to mitigate internal privacy threats. First of all, privacy concerns of OSN users are investigated by taking the most famous OSN site, Facebook, as a case study. Secondly, friendship intensity calculation metrics are explored to use with data mining techniques in order to predict friendship levels. Furthermore, this study purposes a privacy preservation framework which utilizes friendship levels information.

This study has identified several crucial aspects of privacy in OSNs. Although, privacy preferences of individuals are little diverse but, almost every OSN user is concerned about his/her privacy. There is no opinion difference in terms of their nationality, gender and Internet activity level regarding privacy preservation. The OSN users with small sized social network are relatively more concerned about their privacy. Furthermore, several users tend to hide their most private information as a precaution to preserve their private data. This pattern of information hiding varies with respect to the gender. We also found that most of the individuals who are hiding their information are capable of using Facebook privacy settings. This aspect shows sort of dissatisfaction over the privacy setting provided by the social network service provider.

Before exposing in the real life, we assess numerous things such as sensitivity of information, possible gains from the information revelation and, vulnerability of environment. This kind of careful attitude is also required to maintain privacy in OSNs. Unlike the previous studies, a good number of Facebook users change default privacy settings which lead to the fact that OSN users are getting more conscious about the privacy issues. Moreover, the change in privacy settings also depends on the user activity level on Internet and Facebook. This study found little contradictory behavior of OSN users regarding their privacy. A number of individuals keep on using some particular OSN facility which they also criticize as a privacy threat. Furthermore, users are concerned regarding the privacy threats from their own network. That is why; many users only want to expose their data to a certain number of their friends. These factors lead to the subsequent objective of this research.

The main objective of this thesis is to identify methods and metrics for friendship intensity/level prediction. This thesis has suggested several improvements in the interaction count method. In this work, we found individuals prefer one interaction type over the other while communicating in OSNs. Therefore, simple accrual of these interactions reduces relative importance of some specific interaction type. Furthermore, users' activity level on OSN also affects their interaction behavior with their friends. This research also identifies ways to integrate these factors to enhance the performance of interaction count method. Furthermore, some studies have shown that people only maintain interactions with a small number of their friends in OSNs [4, 50]. Therefore, interaction based methods may only provide correct intensity for a certain subset of one's friends. In order to deal with this overhead, this study emphasizes to explore non interaction based features for friendship intensity calculation. Mutual friends and total profile visits are identified as useful non interaction based metrics. We also establish that data mining classification techniques are quite suitable for predicting the friendship levels.

5.2 Future work

Friendship intensity calculation is a first step in the process to improve the privacy on OSNs. This study introduces and demonstrates the use of data mining for friendship

intensity calculation. The experimentation of this study is based on the artificial data which is developed mostly through online survey. This data mining model is not validated on real data of Facebook. Although, we recommend this functionality as vendor level solution however, we are planning to develop a Facebook application based on our data mining framework in order to validate its performance. One further direction of this work is to explore the applicability of this framework on other OSNs such as MySpace and Twitter.

Other good direction could be to look into the ways to find the "quality of interaction" before predicting the levels of the relationship. This information could be crucial for friendship intensity calculation function in order to improve its accuracy. Therefore, another important direction is to explore the techniques that utilize NLP and Text Mining for classifying interactions [48].

Interactions based metrics are much important to estimate the quality of relationship between individuals but we cannot entirely rely on these metrics because individuals only maintain interactions with certain number of their friends [50]. We already took a stride in this direction by identifying two non-interaction based methods. We are also planning to experiment these methods separately or by integrating them with interaction based methods. We are also interested in comparing these metrics with interaction based metrics.

6 REFERENCES

[1] D. Boyd, "Social Network Sites: Definition, History, and Scholarship," *Journal of computer-mediated communication*, vol. 13, pp. 210-230, 2007.

[2] Wikipedia, "List of social network websites," ed. http://en.wikipedia.org/wiki/List_of_social_networking_websites, Retrieved on 8 March, 2010.

[3] L. A. Cutillo, *et al.*, "Safebook: a privacy-preserving online social network leveraging on real-life trust," *IEEE Communications Magazine*, vol. 47, pp. 94-101, 2009.

[4] L. Banks and S. F. Wu, "All friends are not created equal: an interaction intensity based approach to privacy in online social networks," in *2009 International Conference on Computational Science and Engineering (CSE), 29-31 Aug. 2009*, Piscataway, NJ, USA, 2009, pp. 970-4.

[5] C. Wohlin, *Experimentation in software engineering : an introduction*. Boston: Kluwer, 2000.

[6] J. W. Creswell, *Research design : qualitative, quantitative, and mixed methods approaches*, 2. ed. Thousand Oaks: Sage, 2003.

[7] E. A. Baatarjav, *et al.*, "BBN-based privacy management system for Facebook," in *2009 IEEE International Conference on Intelligence and Security Informatics (ISI), 8-11 June 2009*, Piscataway, NJ, USA, 2009, pp. 194-6.

[8] A. Acquisti and R. Gross, "Imagined communities: Awareness, information sharing, and privacy on the facebook," in *6th International Workshop on Privacy Enhancing Technologies, PET 2006, June 28, 2006 - June 30, 2006*, Cambridge, United kingdom, 2006, pp. 36-58.

[9] A. Acquisti and R. Gross, "Imagined communities: awareness, information sharing, and privacy on the Facebook," in *Privacy Enhancing Technologies. 6th International Workshop, PET 2006. Revised Selected Papers, 28-30 June 2006*, Berlin, Germany, 2006, pp. 36-58.

[10] D. Boyd, "Implications of user choice: The cultural logic of my space or facebook?," *Interactions*, vol. 16, pp. 33-36, 2009.

[11] I. H. Witten, "Data Mining : Practical Machine Learning Tools and Techniques," 2005.

[12] S. Milgram, "The small world problem," *Psychology Today*, vol. 2, p. 60, 1967.

[13] Wikipedia, " Social network service," ed. http://en.wikipedia.org/wiki/Social_Networking_Sites, Retrieved on March 8, 2010.

[14] M. S. Granovetter, "The strength of weak ties," *The American journal of sociology*, vol. 78, p. 1360, 1973.

[15] Musiał and Katarzyna, "RECOMMENDATION SYSTEM FOR ONLINE SOCIALNETWORK," ed: Blekinge Institue of TechnologyMaster ThesisSoftware EngineeringThesis no: MSE-2006:11, July 2006.

[16] L. Garton, *et al.*, "Studying Online Social Networks," Haythornthwaite, *et al.*, Eds., ed: Journal of Computer-Mediated Communication, June 1997.

[17] N. B. Ellison, "FEATURE Social network sites and society: current trends and future possibilities," *Interactions*, vol. 16, p. 6, 2009.

[18] R. A. Hanneman and M. Riddle, "Introduction to social network methods," ed. Internet <http://faculty.ucr.edu/~hanneman/nettext/> (06.03.2010): Online Book, 2005.

[19] S. Wasserman, "Social network analysis: Methods and applications," 1994.

[20] B. Howard, "Analyzing online social networks," *Communications of the ACM*, vol. 51, pp. 14-16, 2008.

[21] A. Mislove, "Measurement and analysis of online social networks

Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC 07 IMC 07," p. 29, 2007.

[22] J. Leskovec, *et al.*, "Microscopic evolution of social networks," in *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008, August 24, 2008 - August 27, 2008*, Las Vegas, NV, United states, 2008, pp. 462-470.

[23] R. Kumar, "Structure and evolution of online social networks
Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 06 KDD 06," p. 611, 2006.

[24] S. B. Barnes, "A privacy paradox: Social networking in the United States," *First Monday*, vol. 11, 2006.

[25] B. Danah, "Making Sense of Privacy and Publicity," ed. SXSW. Austin, Texas, March 13.

[26] R. Gross, *et al.*, "Information revelation and privacy in online social networks," in *WPES'05: 2005 ACM Workshop on Privacy in the Electronic Society, November 7, 2005 - November 7, 2005*, Alexandria, VA, United states, 2005, pp. 71-80.

[27] B. Krishnamurthy, "Characterizing privacy in online social networks
Proceedings of the first workshop on Online social networks - WOSP 08 WOSP 08," p. 37, 2008.

[28] C. Xi and S. Shuo, "A literature review of privacy research on social network sites," in *2009 International Conference on Multimedia Information Networking and Security (MINES 2009), 17-20 Nov. 2009*, Piscataway, NJ, USA, 2009, pp. 93-7.

[29] P. Pecho and J. Nagy, "Social networks security," in *2009 Third International Conference on Emerging Security Information, Systems and Technologies (SECURWARE), 18-23 June 2009*, Piscataway, NJ, USA, 2009, pp. 321-5.

[30] L. A. Cutillo, *et al.*, "Privacy preserving social networking through decentralization," in *2009 Sixth International Conference on Wireless On-demand Network Systems and Services (WONS 2009), 2-4 Feb. 2009*, Piscataway, NJ, USA, 2009, pp. 142-52.

[31] L. Banks, *et al.*, "Davis social links: leveraging social networks for future Internet communication," in *2009 Ninth Annual International Symposium on Applications and the Internet (SAINT 2009), 20-24 July 2009*, Piscataway, NJ, USA, 2009, pp. 165-8.

[32] L. Kun and E. Terzi, "A framework for computing the privacy scores of users in online social networks," in *2009 Ninth IEEE International Conference on Data Mining (ICDM 2009), 6-9 Dec. 2009*, Piscataway, NJ, USA, 2009, pp. 288-97.

[33] K. Juszczyszyn, *et al.*, "Temporal changes in connection patterns of an email-based social network," in *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 9-12 Dec. 2008*, Piscataway, NJ, USA, 2008, pp. 9-12.

[34] W. Barry, *et al.*, "Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community," vol. Annu. Rev. Sociol. 1996 22:213-38, ed: Annual Review of Sociology, 1996, pp. 213-238.

[35] W. B. Hansen, *et al.*, "Network Genie User's Manual," ed. Available at: https://secure.networkgenie.com/admin/documentation/Network_Genie_Manual.pdf: Tanglewood Research, Inc., Retrieved on March, 2010.

[36] A. Mislove, *et al.*, "Measurement and analysis of online social networks," in *IMC'07: 2007 7th ACM SIGCOMM Internet Measurement Conference, October 24, 2007 - October 26, 2007*, San Diego, CA, United states, 2007, pp. 29-42.

[37] C. L. Chang and C.-L. Chang, "Browsing newsgroups with a social network analyzer Proceedings Sixth International Conference on Information Visualisation IV-02," p. 750, 2002.

[38] A. Internet, "Top Sites," ed. http://www.alexa.com/topsites, Retrieved on March 8, 2010.

[39] Wikipedia, "Facebook," ed. http://en.wikipedia.org/wiki/Facebook, Retrieved on April, 2010.

[40] M. Eirinaki and M. Vazirgiannis, "Web mining for Web personalization," *ACM Transactions on Internet Technology*, vol. 3, pp. 1-27, 2003.
[41] T. N. Jagatic, "Social phishing," *Communications of the ACM*, vol. 50, p. 94, 2007.
[42] K. Gammon, " Networking: Four ways to reinvent the Internet," ed. Available at, http://www.nature.com/news/2010/100203/full/463602a.html: Nature, Published online 3 February 2010.
[43] C.-m. Au Yeung, *et al.*, "Decentralization: The Future of Online Social Networking," ed, 2009.
[44] T. Tran, *et al.*, "Design and implementation of davis social links OSN Kernel," in *4th International Conference on Wireless Algorithms, Systems, and Applications, WASA 2009, August 16, 2009 - August 18, 2009*, Boston, MA, United states, 2009, pp. 527-540.
[45] D. J. Solove, "The end of privacy? [social networking]," *Scientific American (International Edition)*, vol. 299, pp. 79-83, 2008.
[46] E. Aimeur, *et al.*, "UPP: user privacy policy for social networking sites," in *2009 Fourth International Conference on Internet and Web Applications and Services (ICIW 2009), 24-28 May 2009*, Piscataway, NJ, USA, 2009, pp. 267-72.
[47] R. Feldman and Mihalcea, "The Text Mining HandbookÃ¢Â Â“Advanced Approaches in Analyzing Unstructured Data," *Computational linguistics*, vol. 34, p. 125, 2008.
[48] A. Kao, "Natural Language Processing and Text Mining," 2006.
[49] T. M. Mitchell, *Machine learning*. New York: McGraw-Hill, 1997.
[50] C. Wilson, "User interactions in social networks and their implications Proceedings of the fourth ACM european conference on Computer systems - EuroSys 09 EuroSys 09," p. 205, 2009.

7 APPENDIX

7.1 Table of acronyms

BTH	Blekinge Institute of Technology
CMC	Computer Mediated communication
DSSN	Device Supported Social Networks
OSNs	Online Social networking sites
RQ	Research Question
SC	Social Capital
SNA	Social Network Analysis
WEKA	Waikato Environment for Knowledge Analysis
DoS	Denial of Services
DDoS	Distributed Denial of Services
SQL	Structured Query Language
CV	Cross Validation

7.2 Privacy survey questionnaire

Survey on User Privacy Concerns Regarding Facebook

Privacy deals with the users' ability to reveal themselves or their information selectively[14]. Privacy preferences of individuals regarding their private data are not same rather it varies in terms of information type as well as the relationship quality. The social networking websites (SNS) such as Facebook can capture the information about each aspect of its users. This information about Facebook users may include personal information (e.g. name, email, phone number, home address, office address etc.), pictures, videos, interests, affiliations (e.g. friends list, joined groups etc.), activities (e.g. sending or accepting friendships request, scores in a game, sending gifts , status updates etc.), comments and lot more. **The use and misuse of this information without your (as a Facebook user) consent, is violation of your privacy.** These violations could be either internal or external. This survey is mainly related to the internal privacy threats which come from your own social network. In Facebook perspective, your social network consists of your friends, third party social games or applications (you normally grant the access of your personal information before using these application), and joined groups or communities. We are only concerned with the privacy threats which are posed by your friends and your awareness about this particular issue.

This survey is only meant for Facebook users. We request for your active and honest participation in this survey in order to achieve our best efforts towards privacy preserving social networks. Thanks for your anticipation!

Background Questions

1. What is your nationality?
2. What is your age?
- Less than 20
- Between 20 to 40
- More than 40

3. What is your gender?
- Male
- Female

4. Your experience of using computers and internet?
- Expert
- Good
- Average
- Beginner

Survey Questions

5. How do you consider yourself to be on Facebook?
- **Very Active** : I share content, use applications and games, update my current status and interact with friends
- **Active :** I mostly interact with friends and rarely use games and application
- **Rare :** I only respond to the alerts
- **Very Rare:** I only use when it is required

6. How many friends do you have on Facebook?
- Less than 50

[14] http://en.wikipedia.org/wiki/Privacy

- 50 to 100
- 100 to 200
- 200 to 500
- 500 to 1000
- More than 1000

7. Do you add people you don't know in real life as friends on Facebook?
   - Yes
   - Only if they have common friends with me
   - No
8. Are you concerned about protecting your private data i.e. phone, email, private pictures etc. on Facebook from any kind of misuse?
   - Yes
   - No
   - Not Sure
9. Do you hide yourself (not showing your complete data) on Facebook due to the privacy concerns?
   - Yes
   - No
10. Do you submit false personal information to Internet services in general due to privacy concerns?
   - Yes
   - No
   - Sometimes
11. Do you feel that any of your Facebook friends could be a threat to your privacy?
   - Yes
   - No
   - Maybe
12. Are you willing to share your most private information with?
   - Only some of the friends in your friend's network
   - All friends in your friend's network
   - Friends of Friends
   - Everyone
13. In your opinion, your private data is more at risk from?
   - Your Internal Network (from your friends)
   - Your friends of friends
   - Third party Facebook applications

Facebook Privacy Settings

14. Do you change Facebook default privacy settings?
   - Yes
   - No
   - I know they exist but never change it
   - I don't know whether they exist or not
15. According to you, how understandable are the privacy settings?
   - Easy
   - Difficult
   - Very Difficult
16. Do you know exactly, how much of you private data others can see without being your friend?
   - I don't Know
   - I Know
17. Do you think that there should be some automatic or semi-automatic privacy preservation mechanism?

- Yes
- No

18. Government agencies can examine Facebook data. According to Facebook privacy policy; "We may also share information when we have a good faith belief it is necessary to prevent fraud or other illegal activity, to prevent imminent bodily harm, or to protect ourselves and you from people violating our Statement of Rights and Responsibilities. **This may include sharing information with other companies, lawyers, courts or other government entities**."[15] In real life government agencies may not enter into your private life without proper court orders. How do you comment on this provision?
    - I strongly condemn this
    - I strongly support this
    - I think it has no harm

Your communication behavior with your Good Friends*

* Your G**ood friends** are those who will not pose any privacy threat to you. Just imagine those friends in your friend's network; you are sure that they will never misuse your personal information. You can call them your good friends in this scenario. They could be some of your classmates, colleagues or relatives. They could also be your parents, siblings or spouse. In other words, you are ready to share almost all of your information with your good friends and you are certain that they will not misuse that information.

19. How do you interact with your good friends on Facebook, mostly?
    - Private Messaging
    - Wall Posting
    - Commenting the posts
    - Chatting
    - Picture Tagging
    - Sending Gifts
    - Wishing
    - Playing Games
20. Approximately, how many times you interact with your good friends on Facebook by using any of the interaction method described above?
    - Once in a week
    - Many times in a week
    - Once in month
    - Many times in month
    - Once in a year
    - Many times in a year
21. Do you like to visit the profiles of your good friends, often?
    - Yes
    - No

[15] http://www.facebook.com/policy.php

7.3 Email of survey invitation

Subject: Facebook Privacy Survey

Dear All,

According to Facebook,

"If Facebook were a country then it would have been a third largest country of the World with a population of 400 million People"

In spite of being mostly a free service, Facebook's net worth is more than 5 billion US dollars which makes it one of top corporate of the world, why! It is because of tremendous amount of user data, in form of names, addresses, phone numbers, comments, videos, pictures etc. Will Facebook be really able to preserve the privacy of its enormous users? In Blekinge Institute of Technology (BTH), we are conducting a research to explore techniques that can be used to improve privacy of social network sites. We need your contribution in this research effort by just filling the following completely anonymous online survey.

http://www.mysurveylab.com/index.php?cId=ebcbbef53de8ef39dc2dc3726eae3f98c24873a8&pid=133&lng=en

This survey will also enhance your awareness about privacy related threats in social network sites. We will really appreciate if you spare just 10 minutes for this online questionnaire.

For further information please email at: bth.fb.survey@gmail.com
For discussion please join following FB group:
http://www.facebook.com/group.php?v=wall&ref=nf&gid=110732835610236

Regards,

7.4 Training Data Set

Messages	Commenting	Chatting	Wall_Posting	Tagging	Profile_Visit	Level
9	11	2	24	7	1	good
30	7	8	29	1	0	very_good
16	16	25	17	4	1	very_good
17	4	11	12	2	1	good
17	25	29	7	13	0	very_good
29	6	23	29	0	0	very_good
20	29	16	26	13	1	very_good
0	13	5	24	0	1	good
19	18	20	28	0	1	very_good
22	16	22	18	11	1	very_good
30	20	27	19	9	1	very_good
30	8	28	12	2	1	very_good
1	5	10	6	0	0	average
23	17	31	10	7	1	very_good
6	5	17	27	11	0	good
18	22	0	13	5	1	good
11	9	31	7	10	0	good
3	23	11	13	13	0	good
26	20	29	24	0	0	very_good
26	8	20	10	2	0	good
21	22	5	20	2	1	very_good
11	16	28	9	11	1	very_good
15	27	13	24	4	0	very_good
5	5	3	3	13	0	average
18	23	13	6	1	1	good
21	23	11	24	6	1	very_good
12	23	3	8	1	1	good
29	17	26	9	5	0	very_good
23	16	11	24	9	0	very_good
32	6	3	12	6	0	good
0	2	0	11	12	0	average
31	8	8	1	7	0	good
31	13	29	23	10	0	very_good
32	14	28	10	4	1	very_good
22	8	25	24	9	0	very_good
1	21	9	1	1	0	average
17	1	14	23	1	0	good
4	1	14	2	12	0	average
12	25	23	8	2	0	good
27	26	11	18	9	0	very_good
27	22	24	4	11	1	very_good
10	10	29	23	1	0	very_good
26	8	2	7	4	1	good
16	21	25	7	0	0	good
10	25	28	20	6	1	very_good
24	9	10	15	6	0	good
13	23	14	10	7	1	very_good

11	10	27	13	8	0	good
13	3	29	6	5	1	good
7	20	25	3	4	0	good
22	2	18	0	8	0	good
4	0	25	18	4	0	good
6	21	22	1	2	0	good
7	17	13	4	11	1	good
3	0	14	26	1	1	good
4	23	28	21	11	1	very_good
23	28	16	15	7	1	very_good
28	21	18	16	10	1	very_good
28	0	10	8	2	0	good
1	8	3	15	2	0	average
10	9	0	6	0	1	average
23	21	23	7	3	0	very_good
3	2	31	20	8	1	good
13	14	24	22	10	1	very_good
27	24	6	28	4	0	very_good
5	18	22	4	11	1	good
31	16	13	20	11	1	very_good
31	26	24	28	7	0	very_good
23	29	28	14	10	1	very_good
14	16	14	19	7	1	very_good
1	20	12	21	11	0	good
13	24	25	25	6	1	very_good
17	14	15	8	6	0	good
22	4	0	17	9	1	good
7	2	10	20	10	1	good
15	23	16	26	6	0	very_good
3	27	4	27	10	1	very_good
16	8	3	28	8	1	good
0	7	22	0	11	0	average
25	2	14	2	10	0	good
30	7	30	13	1	0	very_good
25	6	10	27	1	0	good
18	18	12	20	8	1	very_good
7	17	7	27	8	0	good
9	12	5	21	7	1	good
29	13	5	25	12	0	very_good
9	19	17	18	8	1	very_good
23	4	9	3	11	1	good
2	1	16	9	13	0	average
13	8	17	16	0	0	good
16	27	15	20	10	0	very_good
32	0	31	15	8	1	very_good
6	10	23	15	6	1	good
13	16	19	27	12	1	very_good
1	8	28	16	11	1	good
19	10	25	2	0	1	good
13	0	18	12	7	0	good

32	21	18	21	3	0	very_good
8	19	19	26	7	1	very_good
15	17	25	10	1	1	very_good
8	22	31	1	11	0	very_good
19	7	1	4	9	1	average
12	19	9	28	5	1	very_good
0	18	5	19	7	1	good
4	29	9	28	4	0	very_good
19	17	5	17	8	1	good
19	6	7	19	5	1	good
7	22	16	16	9	1	very_good
2	21	16	7	4	0	good
32	14	0	7	6	0	good
22	0	30	26	8	1	very_good
28	5	26	2	11	0	very_good
31	1	30	14	8	1	very_good
12	19	12	26	12	1	very_good
30	26	7	28	2	1	very_good
9	25	22	7	10	0	very_good
27	17	15	19	9	0	very_good
28	18	11	24	1	1	very_good
15	14	19	26	5	1	very_good
20	9	13	25	3	1	very_good
17	16	14	7	11	0	good
17	19	23	29	4	0	very_good
10	4	0	20	13	0	good
20	13	20	7	4	0	good
25	27	11	6	11	1	very_good
26	9	5	7	7	1	good
13	2	19	4	1	1	average
17	11	2	17	0	0	good
23	20	14	10	12	1	very_good
12	21	22	22	12	0	very_good
11	21	26	16	13	0	very_good
16	9	14	3	6	1	good
1	0	29	3	8	1	average
4	8	21	17	8	1	good
12	0	9	2	10	0	average
19	4	28	12	9	1	very_good
26	5	5	7	4	1	good
26	10	20	21	6	0	very_good
23	17	16	11	2	0	good
32	16	6	7	8	1	very_good
15	6	3	11	8	1	good
6	12	30	11	9	0	good
0	14	19	20	3	0	good
30	27	19	13	4	1	very_good
19	8	11	23	7	1	very_good
25	0	6	26	5	1	good
21	17	9	18	13	1	very_good

28	29	14	14	9	1	very_good
17	17	19	4	0	1	good
31	18	5	8	5	1	very_good
9	4	23	14	0	0	good
11	20	22	18	4	1	very_good
19	5	20	1	9	0	good
23	5	28	9	4	1	very_good
7	8	16	12	7	1	good
4	25	2	26	4	0	good
32	19	25	20	3	0	very_good
11	25	8	0	12	1	good
10	19	3	6	13	0	good
3	5	23	18	8	0	good
18	22	4	18	8	0	good
24	19	4	16	4	0	good
21	20	8	27	9	0	very_good
28	0	26	29	3	1	very_good
13	10	31	4	13	1	very_good
14	16	11	17	2	1	good
2	27	20	5	6	1	good
20	3	20	26	0	0	good
2	26	12	27	12	1	very_good
22	8	31	20	6	1	very_good
8	26	17	18	6	1	very_good
19	16	19	6	10	0	good
6	10	5	25	3	0	good
6	9	3	17	3	1	average
6	27	7	6	8	1	good
6	25	17	22	6	0	very_good
14	2	17	1	4	0	average
4	14	20	9	4	1	good
16	15	5	18	1	0	good
18	9	9	24	7	1	very_good
15	21	23	1	6	0	good
2	27	9	18	13	1	very_good
30	2	2	16	5	0	good
12	27	30	16	7	0	very_good
31	2	12	7	12	1	good
4	11	20	26	6	0	good
3	14	0	11	1	1	average
32	0	2	1	7	0	average
15	28	11	13	7	1	very_good
16	13	31	26	9	1	very_good
26	25	9	13	10	1	very_good
18	24	6	17	3	0	good
18	5	11	18	3	1	good
3	22	27	13	2	0	good
11	20	26	10	7	1	very_good
8	11	11	16	11	0	good
8	11	29	14	13	0	very_good

26	15	10	4	2	1	good
1	7	10	26	9	1	good
8	2	6	2	5	1	average
3	6	4	1	4	0	low
5	0	3	7	1	0	low
2	6	2	5	0	0	low
0	7	0	6	4	1	average
1	6	7	0	4	1	average
0	2	5	4	2	1	low
5	0	4	0	4	1	low
2	1	5	5	4	1	average
0	6	5	1	2	1	low
0	5	3	8	3	0	low
7	5	5	4	1	0	average
0	7	1	6	4	0	low
5	2	2	2	4	0	low
0	0	7	6	0	0	low
3	1	0	8	0	1	low
2	7	3	8	2	1	average
8	4	4	1	1	1	average
1	6	6	4	0	1	average
4	0	6	3	4	1	average
7	2	1	8	0	0	low
6	3	2	1	1	1	low
7	2	2	3	2	1	low
8	4	3	6	0	1	average
4	5	5	2	0	1	average
1	7	6	3	4	1	low
5	4	2	8	3	1	low
6	4	7	0	4	0	low
5	6	0	1	2	0	low
6	0	5	3	1	1	low
7	3	5	5	4	1	average
5	3	1	8	2	1	average
4	0	3	4	0	0	low
7	6	1	7	4	1	average
8	4	5	6	4	0	average
4	6	4	2	2	0	low
7	0	6	0	4	0	low
8	3	2	2	4	0	low
5	6	1	6	3	0	average
3	5	5	1	1	1	low
1	7	4	6	0	0	low
0	7	0	8	4	1	average
1	7	3	5	0	0	low
0	6	2	2	2	1	low
3	7	5	1	0	0	low
7	2	7	3	0	0	low
3	0	0	4	0	0	low
4	2	7	8	4	0	average

3	7	6	5	0	0	average
5	4	1	2	4	0	low
3	5	4	2	0	1	low
3	3	5	3	4	1	average
3	3	0	3	1	1	low
4	6	6	2	2	1	average
5	0	4	7	1	0	low
0	5	0	5	0	0	very_low
1	3	5	7	3	1	average
5	1	0	5	2	0	low
4	0	0	2	1	1	low
2	7	4	7	4	1	average
2	7	5	2	2	1	average
3	5	0	0	2	1	low
5	1	1	5	3	1	low
0	2	6	1	2	0	low
6	6	1	2	3	1	average
2	7	6	8	1	0	average
5	5	0	7	3	1	average
6	3	6	7	0	1	average
5	0	6	3	1	0	low
5	7	2	6	0	1	average
6	3	6	5	1	0	average
8	7	5	8	3	0	average
5	3	6	3	1	1	average
8	7	7	6	1	0	average
1	7	0	4	3	0	low
3	3	5	2	4	1	average
3	6	7	1	4	1	average
5	3	1	3	4	1	low
3	6	3	2	2	1	low
4	5	5	5	1	0	low
8	5	5	5	4	1	average
4	2	4	4	2	0	low
2	0	6	4	2	1	low
8	6	3	3	0	1	average
1	5	5	8	1	1	average
2	1	4	0	3	1	low
6	2	7	2	4	0	average
5	0	4	4	2	0	low
1	4	6	5	4	0	low
5	1	5	3	0	0	low
2	3	2	8	2	0	low
4	1	2	4	1	0	low
8	0	2	3	2	1	low
3	5	3	2	1	0	low
3	0	5	0	1	1	low
1	5	1	3	3	1	low
3	1	6	7	3	0	low
0	4	1	3	4	1	low

2	3	3	6	3	1	average
2	0	6	8	0	0	low
6	1	7	8	3	1	average
5	0	5	3	4	0	low
4	6	7	5	1	1	average
3	4	2	7	1	1	average
5	4	7	4	2	0	average
7	2	2	6	1	1	average
3	6	7	0	0	1	low
4	5	5	3	1	1	average
2	5	3	4	2	1	low
4	1	1	4	3	1	low
5	3	0	7	1	0	low
6	4	0	6	3	0	low
3	1	7	0	2	0	low
8	6	7	3	1	0	average
0	7	4	2	2	0	low
0	7	6	0	4	0	low
2	4	4	3	1	0	low
5	5	3	5	3	0	average
7	5	6	7	0	0	average
2	0	1	8	1	1	low