Inferring Private Attributes in Online Social Networks

Nasireddin Shadravan

Master of Science Thesis

Network Architectures and Services Group (NAS)
Department of Electrical Engineering, Mathematics and Computer Science
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

TU Delft Delft University of Technology

Network Architectures and Services Group

Inferring Private Attributes in Online Social Networks

Master of Science Thesis

For the degree of Master of Science in
Network Architectures and Services Group (NAS)
at Department of Electrical Engineering, Mathematics and Computer Science
at Delft University of Technology

Nasireddin Shadravan

August 13, 2012

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Delft, The Netherlands

TUDelft Delft University of Technology

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Delft, The Netherlands

Abstract

Online social networks (OSNs) are playing an important role in current world and the way people communicate with each other. Despite the advantage of using online social networks, there are certain privacy risks that can affect users of such services. Since users provide a lot of personal information in OSNs, concerns about how data placed in online social networks may raise among the users. Social networking sites have responded to these concerns by introducing privacy filters to their site, allowing users to specify which aspects of their profile are visible to whom. Such privacy settings is not effectively used by half of the OSN users based on our analysis and we collect large number of public profile information from the well-known social network Hyves.nl in the Netherlands. We then show that public friendship links of a person can expose different attributes about him. Based on friendship links we are able to infer and predict some of the attributes of a user with good accuracy.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would never have been able to finish my thesis without the guidance of my supervisors, help from friends, and support from my family.

I would like to express my deepest gratitude to my advisor, Dr. Christian Doerr, for his guidance, patience, and understanding during doing the thesis. I would also like to thank Ir. Norbert Blenn, who helped me overcome different questions along the way and motivating me on new ideas. Without their help I would have never been able to find the correct direction toward reaching the end of this project.

I would like to thank my brothers and my sister and especially my dear parents who supported my on every part of my life especially when I was so far from them. They have made a huge influence on my life and I would be always grateful to what they have done for me to reach this point.

I would also like to thank Aida Emami, Navid Sarhangnejad and Peyman Shojaei who has truely been my best friends during this time and supported me in every aspect of my life in the past years. Many thanks to them and I hope I can compensate all the support they did for me.

Delft
August 13, 2012

Nasireddin Shadravan

Chapter 1

Introduction

Online social networks (OSNs) have been growing significantly during last few years attracting more than a billion users globally. Without a doubt OSNs have been one of the most influential products in the last decade. They are used for different purposes ranging from daily communication (e.g. Facebook, Hyves, Twitter) to business networking (e.g. LinkedIn, XING). People use OSNs to express their feelings, keep in touch with friends, find new or lost contacts and share common interests with others. Inevitably, OSNs have changed the ways of communication, marketing and even political campaigns where people can organize certain events as a way of expressing their ideas.

While various OSNs have different purposes, most of them share similar feature: Individuals are offered by a "profile" which is a representation of themselves and can be used by others to contact or being contacted. The contact can be finding new friends, communicating with people they already know, business purposes such as finding a new job, receiving recommendations and much more. Based on the purpose of the service, a profile contains various types of information. For example, Facebook is a social network which mainly connects people that already know each other and are trusted. Facebook users provide information about their personal life, photos, life events and so on. On the other hand, in LinkedIn which is a professional service, people state their career life and resume for business networking. Because of the benefits, online social networks are so popular that a high number of social sites are introduced in different topics.

Although OSNs are quite useful in different sense, there has been some considerations about privacy of users in such services. OSNs are large datastores of personal information. This information is valuable in the sense that by statistical analysis it is possible to extract the preference of users based on different criteria such as age and location. Such analysis can then be used for advertising and research purposes. Third-parties provide targeted advertisment to increase their commercial revenue using the social platform and customize their promotions exactly based on the preferences of visitors and increase their chances on marketing. OSN providers such as Facebook state that they will not hand private information to these third-parties. However, there has been many controversies about leakage of sensitive information to third parties where OSN providers handed private user information along with self identifying information. A recent investigation by the Wall Street Journal showed that personal ID of Facebook users was being transmitted to third party advertisement and tracking companies along with their personal interests which was against the promises made by Facebook [1]. This is where concerns are raised about the privacy of OSN users.

The main privacy concern is that members might not be willing to expose their profile information to everyone inside or outside a network. People need control over their personal information and how it is being shown on the web. In OSNs users provide their email address, photos, friends, education, career background, relationship status and activities such as commenting. For various reasons one might be willing to hide them from certain people. Reasons such as safety, separation of work environment and personal life are among them. If the information is public to everyone it can cause problems such as losing a job. Furthermore, it can be collected and used for commercial purposes without the consent of users.

To circumvent the issue, most of OSN providers enable users to customize their settings in order to maintain a level of privacy. These settings allow a user to select the visibility for different parts of one's profile such as photos, relationship status, status updates, hometown and other personal information. This can be helpful and keeps the strangers away from a user's personal information and limits its visibility only to certain people. On the other hand, OSN providers prefer a more public and permissive profile for members in order to engage users in more interactions with each other. Hence, in many cases the default privacy settings in an OSN is not optimal and it needs to be changed by users.

The privacy settings usually does not fully allow hiding friendship links and groups affiliations and the connection between people and groups are publicly visible. Such links and affiliations can lead to information leakage and expose high amount of information. In addition, many users do not protect their profiles from strangers and the network would be a mixture of public and private profiles. As a result, while an individual protects his profile using the privacy settings, it is possible that a large fraction of his friends have an open profile which contains information about him including the friendship link, comments and so on. Also, even if there are no direct information about a person in his friends, by statistical analysis it would be possible to infer some attributes for a user even if he has a private profile which is the topic of this thesis.

The goal of this thesis mainly highlights how it is possible to infer and reconstruct private attributes of OSN users based on friendship links and group affiliations. Using probability models and data mining approaches such as *association rule learning*, it is shown that with certain possibilities it would be feasible to infer private attributes of users. To see the result on a real dataset, a well-known Dutch OSN, Hyves is collected and used. Our analysis shows that it is not easily possible for an active OSN member to fully protect its privacy without keeping a fully private profile.

The rest of this thesis is as follows: Chapter 2 discusses a background about privacy and threats in online social networks. Chapter 3 demonstrates how the Hyves dataset is collected and demographics of the dataset are presented. In chapter 4 we propose our methods to predict and infer user attributes based on friendship link. Chapter 5 shows the evaluation results of our methods and in Chapter 6 conclusions and future works are presented.

Chapter 2

Background

The right of privacy has received legal recognition in many countries and it is evolving by the growth of technologies. Since the wide spread use of online social networks, privacy has become an important topic in OSNs as well. As OSNs play an important role in social interactions, new privacy issues were introduced and researchers have studied different areas such as privacy risks, threats and protection measures in social networks.

In this chapter a definition of privacy and its importance is discussed. Then, we explain how OSNs provide their members with appropriate methods to maintain the privacy. We then discuss the privacy issues in OSNs as the related works.

2-1 Definition and Importance of Privacy

The word “privacy” is an abstract and contentious word which is not easily definable. Roger Clarke [2] defines privacy as *“the interest that individuals have in sustaining a ’personal space’, free from interference by other people and organisations”*. On a deeper level, Clarke extends this definition to several dimensions:

- **Privacy to person:** which is concerned with the integrity of the individual’s body. It relates to physical concerns about a person and includes issues such as blood transfusion without consent and compulsory sterilization.
- **Privacy of personal behavior:** This relates to different aspects of behavior such as sexual preferences, political activities and religious thoughts both in private and public places.
- **Privacy of personal communication:** Individuals have an interest to be able to communicate among each other through different media without being monitored or intercepted by other persons or organisations.

- **Privacy of personal data:** Individuals claim that data about themselves should not be available to other individuals or organisations without their consent and even if the data is processed by a third-party, the individual must be able to have considerable degree of control over it data and its use.

With the growth of online social networks and ease of communication, the last three items are closely linked together. First, users of a social network should not be enforced to private attributes such as relationship status of religious view in order to use the service. Second, many OSN users communicate using services such as photo upload and commenting within the platform. Definitely it would not be pleasant for a user if unknown people or the OSN provider is reading their comments or viewing their photos without permission. Third, when users enter information into an OSN they expect to have control over their content and should be able to remove the content whenever they want.

There are different reasons why privacy is important [2]. Psychologically, people need a private space. Actions take place based on the observations and judgments made by a person and this requires having room for decision making. Also, people need freedom to behave and communicate with others without the concern of being observed or monitored. People need to be free to innovate and should not be enforced to act what they do not prefer. People also need to be free politically. To be free to think, argue and express their opinions without the fear of being convicted [2]. In addition, individuals are members of certain communities in different domains such as work environment, family and friends and they act differently. They often prefer not to intermingle these communities with each other. As an example, a discussion of two close friends might be inappropriate for their families and if disclosed by one of them, it is felt as a violation of one's privacy. Another situation would be when someone tries to get hired by an organization, he might be required to provide information such as religious thoughts which is not relevant to his application but can have an effect on the hiring procedures. If this is publicly available in an online profile it can cause difficulties for the person. However, with the growth of technology the boundaries for secrecy of information are changing and sometimes it is difficult to define a line for it. Google CEO Eric Schmidt once stated: "If you have something that you don't want anyone to know, maybe you should not be doing it in the first place." [3]. This argument brings many counter arguments because privacy is not only about hiding information which are embarrassing. Many people need privacy in order to feel safe and secure. If it is not provided their information can potentially be used later on and might lead to embarrassments for the person.

So far, we discussed about the privacy of information. We need to clarify what type of information is important for maintaining privacy. Many people are not concerned about having their information public unless their identity is linked with it. When information is linked to the identity of a person it can raise concerns and can be used for potential surveillance or unwanted activities. This sensitive and identifying personal information is called *personally identifiable information (PII).* The US National Institute of Standards and Technology (NIST) guidelines define PII as "information which can be used to distinguish or trace an individual's identity, such as the name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc" [4].

When someone's privacy is lost it means that some information containing PII is either collected, analyzed or shared by advertisers, governments or any other third-party without the consent of the owner. In either way, this information can be potentially harmful to that person leading to social embarrassments.

2-2 Privacy Measures Taken by OSN Providers

In early years of online social networks, privacy concerns were not in the center of attention. Service providers such as Friendster and Orkut did not provide any methods for a person to limit the visibility of the content of the profile. In Orkut for example it was possible to view anyone's photos, videos and wall posts. But people started misusing the photos and videos by creating fake profiles. Also viruses were created to attack the OSN users. In an incident a worm was spread across the network as a post on user profiles. Since the profiles were public, any visitor of an infected profile would get infected, username and passwords were hijacked and the worm was propagated into his profile [5]. Such threats made it more important to provide means of maintaining privacy.

As the social platforms grew, many people requested to have a better control on the visibility of their profiles and the content they are posting online. Also OSN providers perceived that privacy needs for users is an important feature needed in their platforms. In order give control of the profile to users, OSN providers define a set of visibility classes which can be assigned to each part of profile and this way, different parts of the profile would be visible only to people in the specified user class. Common predefined user classes are *myself*, *friends*, *friends of friends* and *everyone*. Hence members are able to customize their privacy settings as they prefer. As a result, intended people will see different content comparing to others. Some OSNs such as Facebook allow users to create their own user classes such as family or school friends while others such as Hyves only have the predefined user classes.

A controversial issue in OSNs are the default privacy settings enforced by the OSN provider. Although OSNs provide users with mechanisms to control their content, they prefer a more open platform rather than private communities because more interaction between users leads to expansion of the network [6]. To do so, the OSN provider define a less restrictive privacy settings as the default settings. Since many users are not aware of privacy issues and problems, they trust and accept the default privacy settings recommended by the provider leading to a exposing personal information without the awareness of the user. Matt McKeon has visualized the history of Facebook default privacy settings over 5 years from the beginning [7]. Figure 2-1 shows the difference of default privacy settings in 2005 and April 2010. It is observed that initially in 2005 Facebook default privacy settings were quite restrictive and none of the user information was exposed to the entire Internet (The outer ring). However in 2010 almost all the information was visible to the entire Internet except for birthday and contact information. At some point, Facebook was even accused to make it difficult for a user to change the privacy settings by complicating the process [8].

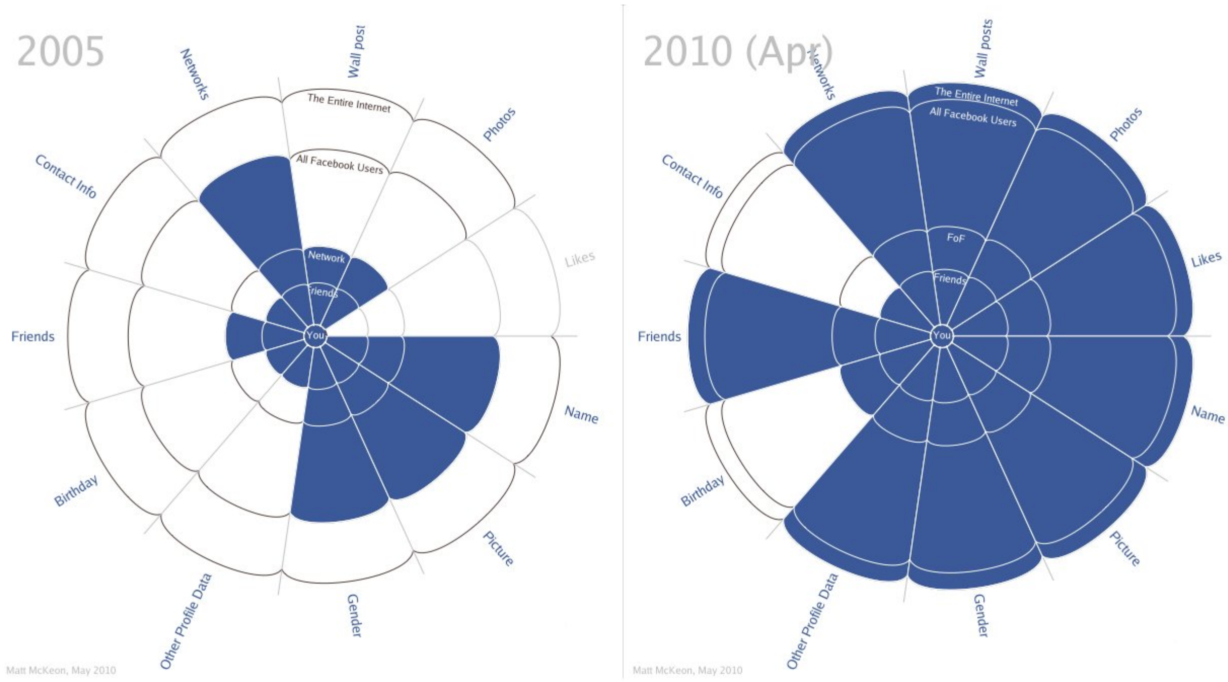


Figure 2-1: Facebook privacy change strategy by Matt McKeon [7]. Blue area have become default public from 2005

As users register in the OSN platform, they start building connections and interacting with others. Some users feel free to post their personal and private data online with the famous quote “I don’t have anything to hide”. A fraction of users that care about their privacy does not use the platform unless they are sure their data is well protected. Another part of the members might be unaware about the privacy controls and how they should be used to hide their profile and since the default settings are not restrictive they might expose their profile unintentionally to strangers. A study in 2005 on a number of CMU students shows that only 0.06% of the students changed their profile visibility [6]. Another study shows that among 67000 twitter profiles, over 99% of them had not changed the default settings in which the tweets, followers and followings are visible to public. For MySpace, out of 3851 random profiles, it was observed that 79% retained the default settings [9]. This suggests that majority of users in OSNs have their information public either unintentionally or find it unimportant.

In summary, although OSN providers enable users to have control on their personal information, but the default privacy settings in many online social networks do not fully protect the privacy unless members explicitly change them [10].

2-3 Related Works

In this section we describe some of the methods used to breach privacy in online social networks. There are two major approaches possible to access private information: passive and active. Active approaches obtain private information by directly attacking the users or service providers while passive methods infer private attributes based on offline statistical analysis of users and their friendship network.

2-3-1 Active Approaches

Active approaches attack targeted or mass number of users directly. A broad range of methods can be used to collect personal user information such as forging profiles, phishing, deception of users and exploiting browser and site vulnerabilities to OSNs.

Forging profiles is a common way of misleading users to trust strangers and obtain their personal information. Since users provide large amounts of personal information in OSNs, an attacker can collect this information and perform a highly personalized and targeted phishing attack. These targeted attacks are called spear phishing attacks. A research study done by Sophos in 2007 shows that out of 200 friend requests sent by a fabricated Facebook profile named Freddi Staur, 87 were confirmed with 82 leaking personal information. In most cases Freddi could have access to family and friends' photos of respondents, information about likes and interests. Moreover, many users provided their family members e.g. parents or siblings [11]. The reason was that default privacy settings does not limit friends from viewing parts of the profile and most of the users had not changed the default settings. Having such information an adversary might forge the identity of a real person and send friendship requests to contacts of that person.

Bilge et al. [12] also automated the process with new attacks: profile cloning and cross-site profile cloning. In profile cloning, the automatic tool *iCloner*, creates a new forged profile of an identifying person with his real name and picture and starts sending friendship request to contacts of that victim. Many users will not get cautious when they know that person even if that person is already in their list. Evaluations showed that the acceptance rate for profile cloning was more than 60%. Cross-site-cloning is different in the sense that it first identifies users that are registered in one social network, but not in the other. Cross-site-cloning has been performed on two business social networks, XING and LinkedIn with a success rate of 56%. The research claims that in these attacks the contacts of the victim may get suspicious and contact the victim about the forged profile. However, in most cases this happens after they have confirmed the friend request.

Exploiting common web vulnerabilities is another active method of obtaining private information of the users. In 2005, Samy Kamkar released the Samy worm in MySpace by exploiting a cross-site scripting vulnerability which allowed him to inject and execute javascript code inside the browsers of other users. The worm spread fast and infected one million users in less than a day. The Samy worm did not collect any personal information but such an attack can easily extract sensitive private information from the profile of users and transmit them to a server controlled by the attacker [13].

2-3-2 Passive Approaches

Passive approaches are mainly based on statistical analyses of users and the friendship network. These passive approaches use an already collected dataset of a social network which can consist of profile information, friendship network, interests of user and membership in groups and communities. The dataset can be collected by fetching the profile information a user specifies, tracking the friendship network through third-party applications, or the combination of different data sources. By having this information, several methods can be used to infer the private attributes based on the information in the dataset.

A study on CMU students using Facebook in 2005 shows that at least 50% of users provide information such as gender, level of study, profile image, birthday, home town, address, phone, high school, relationship status and favorite interests while birthday and profile image was available even for 98% and 90%. Also 89% of users provide their real names comparing to 3% and 8% for only the first name and fake names respectively [6]. Regarding the profile image, 80% of the images contain at least some information useful for identification. Photos uploaded to OSNs can be used by re-identification mechanisms to identify users or identity theft.

Although the OSN providers let the users configure and customize their privacy settings, protecting a single profile or a set of profiles might not preserve the privacy efficiently. In a network which consists of a mixture of public and private profiles it is possible to infer attributes of private users based on information of public users in their network. Mislove *et al.* [14] claim that “you are who you know” because automatic community detection for multiple attributes of users led them infer private attributes with a high accuracy. The research is done over datasets of undergraduate and graduate students of Rice University and New Orleans network. It is observed that users are likely to be friends with those who share their attributes. Moreover, similar users usually tend to form communities centered around attributes. To prove the existence of such communities, the network is divided into communities based on attributes and then strength of these communities is measured. It is observed that these communities exist and are strong. To infer the attributes for the remaining data, two methods are used: global and local. Inferring attributes globally deals with situations such as times when you have the desired attribute for 10% of users and you want to know how well it is possible to infer the other 90%. The result shows that when only 20% of users reveal an attribute such as college or year, it is possible to infer certain attributes of remaining users with an accuracy of 80%. A disadvantage of a global method is that it needs the whole graph data to perform. However, having access to the whole graph is not possible because social graphs are large and only a portion of them can be analyzed. So a local method is preferred to detect the communities. By defining a metric to measure the quality of a community, the algorithm tries to find those subsets of users which yield the highest quality as the result of the algorithm. In addition, an algorithm is also proposed which predicts results given a small number of users with a common attribute.

In addition to basic profile information, it is possible to infer private information from group memberships. OSN providers enable users join different communities such as college or school. In Hyves, members of these communities are visible to everyone even if they have private profiles. Hence, group membership can be used to infer private information about users in a social network. Zheleva *et al.* [15] compared different attacks consisting of friendship links and groups for various social networks. It was observed that the privacy attacks which consider the groups and links together leads to highly accurate predicted results for private attributes. In our study, we aim to infer private attributes based on groups and friendship links. The difference of our work with the above methods is that we use data mining approaches such as *association rule learning* in order to make prediction about users within the network.

Uploading images is a popular feature of many social networks. More than 3 billion photos are uploaded to Facebook each month [16]. In websites like Facebook and Hyves, users can tag their friends inside images with a single click. Considering privacy issues, tagging in photos binds a user to his/her face inside other photos. By collecting enough user photos one can use face recognition mechanisms to identify users across different services. The great improvements in face recognition technologies simplifies such privacy attacks. Currently Face.com has implemented face recognition applications on Facebook enabling users to automatically tag their friends [17]. They make use of the already tagged information of friends and by utilizing learning mechanisms untagged photos are processed and suggestions are made. In a broader point of view any type of identifying information such as image metadata, tags and other information can be used to link user profiles across different social networks for malicious activities.

Pesce et al. [18] exposed some of the privacy issues in photo albums and more specifically in tagging feature introduced by Facebook. They have shown that use of photo tagging feature can enhance accuracy of attacks aiming to predict personal user attributes. The study believes that users that are tagged together in a photo are closer than the average Facebook “friend”. The reasons for that is because tagged friends normally share the same physical location for a reasonable amount of time, tagged friends are slightly more relevant to each other and tagging someone indicates a higher amount of interaction than simply adding him. Using simple algorithms, they were able to predict attributes such as age, gender, hometown and country with accuracy higher than 80%.

Another problem that can lead to privacy issues is sharing information with third-parties or data aggregators. Usually OSN providers give out their information for different purposes such as advertisement, research and statistical analysis. In order to reduce privacy concerns, anonymization algorithms are used to remove the identifiable parts of data such as names, addresses and emails, and then the dataset can be used for market research or other purposes [19], [20]. However, researchers have shown that this method is not fully effective and is vulnerable to certain attacks [19], [21]. Narayanan have proposed a re-identification algorithm to show that a third of users who are verifiable users of both Flickr and Twitter can be recognized in the completely anonymous Twitter graph with only 12% error rate.

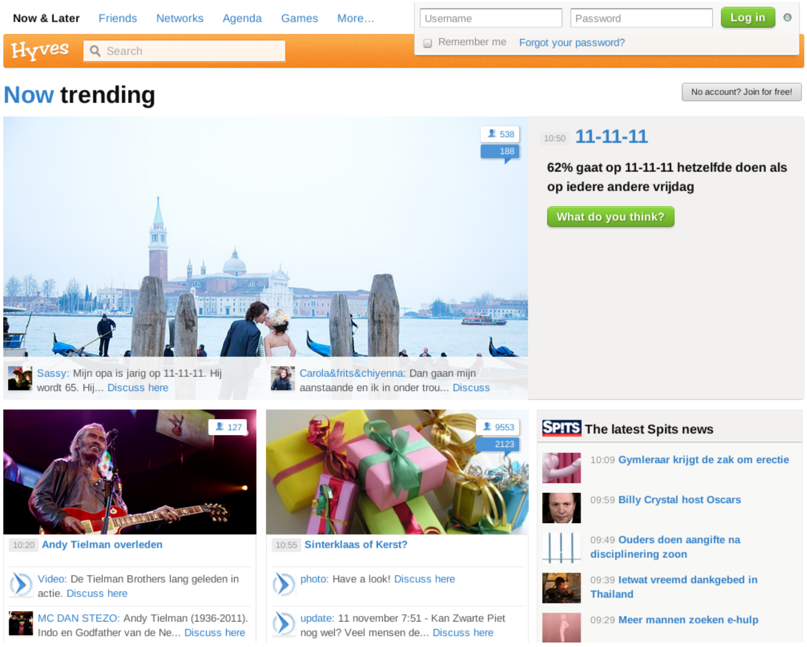
There has been much research to find solutions on how to protect privacy and information in online social networks. Privacy protection can be viewed from the side of an OSN provider or the user. OSN providers claim that the privacy of members is one of their main concerns. In order to share the information with trusted partners, anonymization methods are used to remove sensitive attributes such as real name and user ID from the dataset. The dataset is then shared with third parties for commercial and research purposes. In some cases the dataset was not carefully anonymized and had lead to leakage of sensitive information. Krishnamurthy *et al.* [22] analyzed several social networks and followed the URL patterns made in these networks. In HTTP protocol, when navigating through different pages, HTTP page referrer are sent along with the HTTP request header to other pages. In social networking sites users have a unique identifier and when visiting a profile, this identifier exist in the URL. This is a reference towards a profile containing PII and is used to access the profile information. Also the study has monitored the HTTP requests sent in different social networks and analyzed requests regarding visited profiles, third party applications and advertisements. It was observed that many famous OSN providers including Twitter and Facebook leak user identifiers through HTTP Referrer header. That means alongside the information for marketing, the identity of profile owner was also transmitted to third-party. Hence, the OSN provider could be bypassed and the third-party could send advertisement to users without the intervention of OSN provider. To protect from such attacks, OSN providers should take security and privacy issues more serious and put extra effort on removing PII before redirecting users to third-party services.

Awareness of OSN users plays an important role in protecting their privacy as well. When posting personal information online, one should consider the drawbacks of the action and not rely only on privacy control on the website. Always, there is a possibility of human error which can cause leakage of information. This error can be either from the side of service provider or the person itself. Also, one should not trust every online service. For example, there are enormous number of Facebook applications for different purposes created by third parties. When trying to use any of them, Facebook explicitly informs the users about the information they can get and if a user accepts, he can use the application but his personal information would be available to the application. Users must be skeptical about the applications and should not trust everything unless they are sure about the authenticity of the third-party.

Chapter 3

Hyves.nl

For our analysis we used the social network Hyves.nl. Founded in 2004, Hyves is a widely used online social network in the Netherlands. Based on the statistics provided in the website, as of August 2011 the website had 11.5 million users in a country with 16 million population. Figure 3-1 shows the home page of Hyves.nl.



10:50 11-11-11

62% gaat op 11-11-11 hetzelfde doen als op iedere andere vrijdag

Sassy: Mijn opa is jarig op 11-11-11. Hij wordt 65. Hij... Discuss here

Carola&frits&chiyenna: Dan gaan mijn aanstaande en ik in onder trou... Discuss

10:20 Andy Tielman overleden

Video: De Tielman Brothers lang geleden in actie. Discuss here

MC DAN STEZO: Andy Tielman (1936-2011). Indo en Godfather van de Ne... Discuss here

10:55 Sinterklaas of Kerst?

photo: Have a look! Discuss here

update: 11 november 7:51 - Kan Zwarte Piet nog wel? Veel mensen de... Discuss here

The latest Spits news

10:09 Gymleraar krijgt de zak om erectie

09:59 Billy Crystal host Oscars

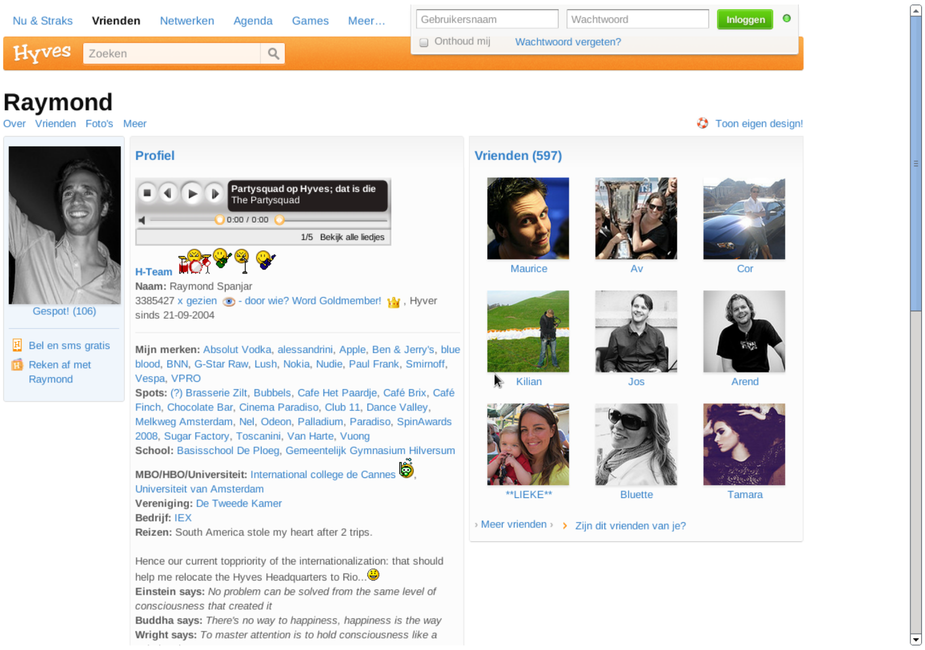
09:49 Ouders doen aangifte na disciplinering zoon

09:39 Ietwat vreemd dankgebed in Thailand

09:29 Meer mannen zoeken e-hulp

Figure 3-1: Home page of Hyves.

Hyves registration is free of charge and there is the possibility to purchase premium accounts to get extra features such as finding out who has visited your profile. Each user has a main profile page which can be accessed by the address *http://username.hyves.nl*. Figure 3-2 illustrates the publicly viewable profile page of Raymond Spanjar, one of the founders of Hyves as an example. The profile page contains the list of some personal information such as name, registration date, a number of friends, number of page visits and interests such as school, music, films. In addition, Hyves allow users to embed their favorite music and videos in their profile page and they can be played within the profile page.



Raymond

Profiel

H-Team

Naam: Raymond Spanjar

3385427 x gezien - door wie? Word Goldmember! , Hyver sinds 21-09-2004

Mijn merken: Absolut Vodka, alessandrini, Apple, Ben & Jerry's, blue blood, BNN, G-Star Raw, Lush, Nokia, Nudie, Paul Frank, Smirnoff, Vespa, VPRO

Spots: (?) Brasserie Zilt, Bubbels, Cafe Het Paardje, Café Brix, Café Finch, Chocolate Bar, Cinema Paradiso, Club 11, Dance Valley, Melkweg Amsterdam, Nel, Odeon, Palladium, Paradiso, SpinAwards 2008, Sugar Factory, Toscanini, Van Harte, Vuong

School: Basisschool De Ploeg, Gemeentelijk Gymnasium Hilversum

MBO/HBO/Universiteit: International college de Cannes, Universiteit van Amsterdam

Vereniging: De Tweede Kamer

Bedrijf: IEX

Reizen: South America stole my heart after 2 trips.

Hence our current toppriority of the internationalization: that should help me relocate the Hyves Headquarters to Rio...

Einstein says: *No problem can be solved from the same level of consciousness that created it*

Buddha says: *There's no way to happiness, happiness is the way*

Wright says: *To master attention is to hold consciousness like a*

Vrienden (597)

Figure 3-2: A user profile in Hyves

More information of the user is available in the "About" page. The About page contains all the personal information provided by the user and is divided into "Overview", "Interests", "Contact" and "Testimonials" sections. The "Overview" section contains information about the person such as relationship status, birthday, living status, schools, colleges and hangouts. Interest section provides different topics that a "Hyver" (users with a registered account) is interested in such as travel, books, music, films and brands. These areas form *groups* and *topics* and are further explained in section 3-2. "Contacts" section includes email and accounts on other social networks if provided and "Testimonials" are descriptions of others about this person.

Users in Hyves, can also upload their photos and share them with friends. It is possible to create photo albums to group the images together. Each of the profiles can also have different view permissions. There are four pre-defined albums in each profile initially:

- **Friends:** Only friends can view the content of the album.
- **Network:** Friends of friends can see the images of this album.
- **Everybody:** Everyone can view this album.
- **Spotted:** Contains the photos other users where the owner is tagged in. A user cannot upload to this album but he can assign permissions to it.

Hyves became popular within two to three years after its launch in 2004 and many people living in the Netherlands joined this social network. Since the registration date is displayed on Hyves profile page, it is possible to see how the registration has emerged through the past few years. Figure 3-3 shows the histogram of user registrations in Hyves. We can see that in 2008, Hyves new users are at its peak but user registration has dropped significantly after that. On one hand, the reason may be that currently a lot of people in the Netherlands have an account on Hyves and the market is saturated. On the other hand, the drop can be due to growth of giant social networks such as Facebook which provide a more international and popular platform in the world having attracted a billion users.

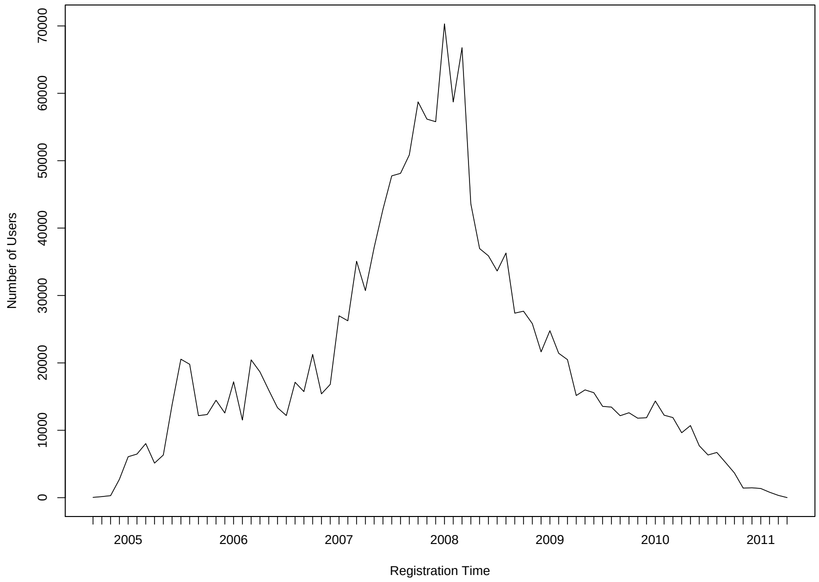


Figure 3-3: Monthly registration of users in Hyves. The y-axis shows the number of users that have registered in Hyves for a specific month.

Registered Hyves users have the ability to determine who can see their profile page. This is achieved by two means: either make the whole profile private or restrict access to specific parts of profile such as name, email, birthday, photos and posts on friend profiles. The access policy has four predefined layers of privacy as depicted in Figure 3-4.

By default, a user profile page is visible to everyone publicly. When the default privacy settings are not changed, profile visibility differs between logged in and logged out state. Without a login, one can see less sensitive attributes of a profile. For example a fraction of interests or first name are visible without a login. In order to see the details of a profile, Hyves asks for registration. Table 3-1 compares the information available when the user is logged in and when logged out. We can see the less sensitive attributes are visible to the entire Internet but for more sensitive data such as hometown, age, friend posts on profile, religion and living status registration is required. If a profile is private to somebody, then the message “Unfortunately this profile is not visible to everyone.” will be shown instead.

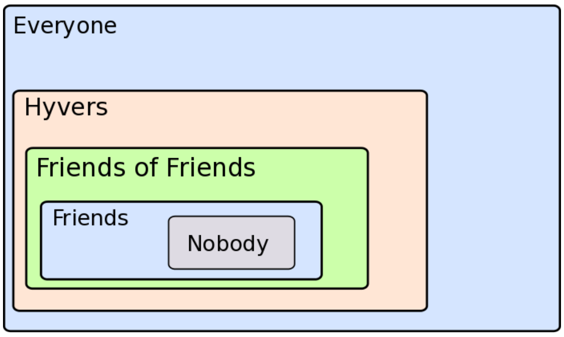


Figure 3-4: Privacy levels in Hyves.

Our examination shows that any attribute that has the possibility of access restriction in privacy settings, is not visible to unregistered visitors. For example, Hyves does not allow users to hide their interests such as favorite brands, music or sports. First name is also an attribute which is publicly available to everyone and no privacy settings exist for hiding that. However, for last name, it is not shown on the profile page without logging in and also users are allowed to select its visibility in privacy settings. It is strange that Hyves has the option to limit the visibility of gender but changing that to any option will not show the gender in the profile.

Additionally, users may join a large selection of groups. Those groups could be real world communities like sport clubs, schools or companies, famous people, bars and restaurants, books, movies etc. Groups lets people communicate based on shared interests similar to a message board. People in groups does not necessarily have social relations but they share common attributes or interests. The size of groups ranges from a few people to several thousands. All the groups in Hyves are classified into different topics. Table 3-2 lists the topics in Hyves with a short description. Each of the topics can be updated by editing the profile and users are automatically added to a group as soon as they list that in their profile. For example, when updating the high school, college, music, brands, sports or any of topics in Table 3-2, the platform searches and suggests from the existing entries and after making a selection user is added to selected group.

Table 3-1: Information available on a non-private profile when a user is logged in and logged out(Default settings).

	Logged In	Logged Out
Name	✔	✔
Age	✔	
Gender		
Hometown	✔	
Living	✔	
Religion	✔	
Contact Information	✔	
Relationship Status	✔	
Registration Date	✔	✔
Page Views	✔	✔
Friends	✔	✔
Friend Posts	✔	
Posts on Profile	✔	
Brands	✔	✔
Hangouts	✔	✔
School	✔	✔
University	✔	✔
Travel	✔	✔
Media	✔	✔
Sport	✔	✔

Table 3-2: List of topics in Hyves

Topic	Description
Brands	Commercial brands and products
Hangouts	Common places where people go such as bars or restaurants
Schools	High schools and lower grades
Colleges	Colleges, undergraduate schools and universities
Clubs	Sports clubs
Companies	Companies
TV Shows	Popular TV programs
Books	Well-known books
Eten	Food
Films	Movies
Games	Computer games
Helden	Celeberities
Media	TV, Radio and Internet channels
Music	Music bands and artists
Reizen	Traveling places
Sports	Sports
Others	Everything which do not fit above

3-1 Demographics

We obtained the data by screen scraping Hyves.nl using multiple parallel breadth first searches. By selecting some random users as seeds, the process collected the profile page, about page, photo tagging information and friendship connections. The process was repeated for newly collected friends. The collected data was then parsed and stored.

Since Hyves blocks high number of repeated requests from a single source in a short time, the crawling procedure was done over a distributed network. Requests were delayed on each node but paralleled among nodes to avoid multiple requests at a time from a single source. Using a central server, new usernames were fed uniformly accross the nodes and after each 100 user profiles were collected, the data was automatically returned back to the central server.

In total, we collected 2,788,487 user profiles. Out of them, 58% are public and the rest are private. By private, we mean that all the attributes are hidden by marking the whole profile visible to only friends or friends of friends. A partially private profile is not considered private because it is not possible to distinguish whether an attribute is private or not provided.

Since the gender is not specified in a profile page, it was not possible to divide the users based on their gender in our dataset. However, according to [23] the percentage of female users in Hyves is much higher than the male users (57% female vs. 43% male). According to “All Facebook Stats” for people living in the Netherlands, the distribution of gender in Facebook is much more equal while the male population is slightly higher (51% male vs. 49% female). That shows women are more dominant in Hyves while it is the other way around for Facebook.

Most users of Hyves are among youth. Figure 3-5 shows the age distribution of users. As illustrated, the majority of users are between 15 and 25 but the overall age range extends to large values. The average age in Hyves is 26 years. Having access to age of users is an important factor in our study. We will discuss different user behaviors based on their age and observation shows that different age ranges have different results in our analysis.

Hyves supports Dutch and English languages. By default, the language is Dutch and it can be switched to English from the bottom of any page. English support of Hyves is not very good as there are Dutch words in some pages with no English substitutes. As we expected and our statistics confirms, most of the members are people living in the Netherlands and Hyves is not actively used in other countries. Since hometown is provided in the profile pages, geographical distribution can be determined. Figure 3-6 shows the cities with highest number of Hyvers. The number of users in the top cities, complies with the population of each city i.e. larger cities such as Amsterdam and Rotterdam have higher number of Hyvers. The largest community of users living outside the Netherlands are Belgians which are less than one percent of the total.

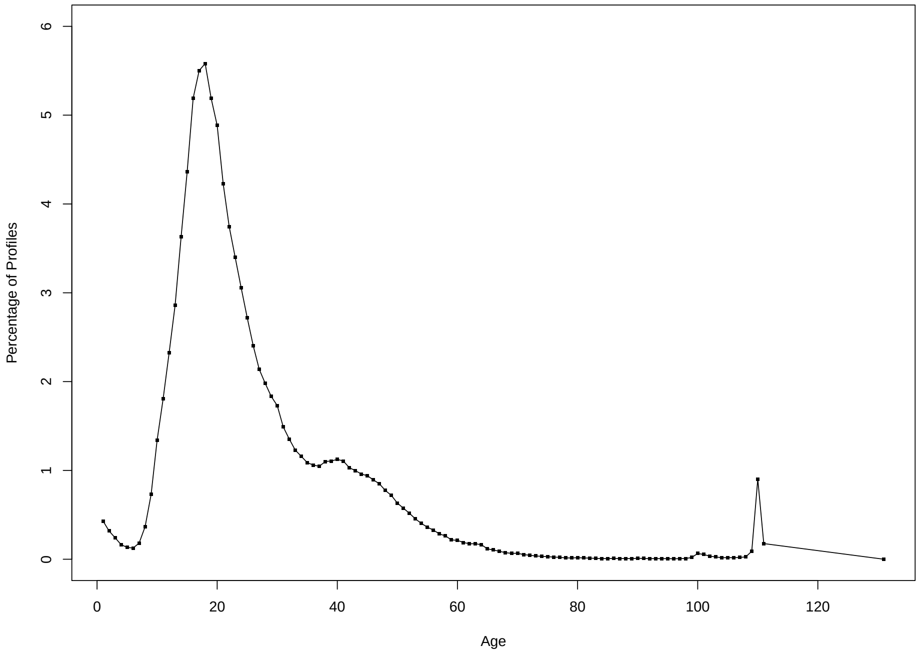


Figure 3-5: Age distribution of Hyves profiles. The majority of users fall into 15-25 age bracket.

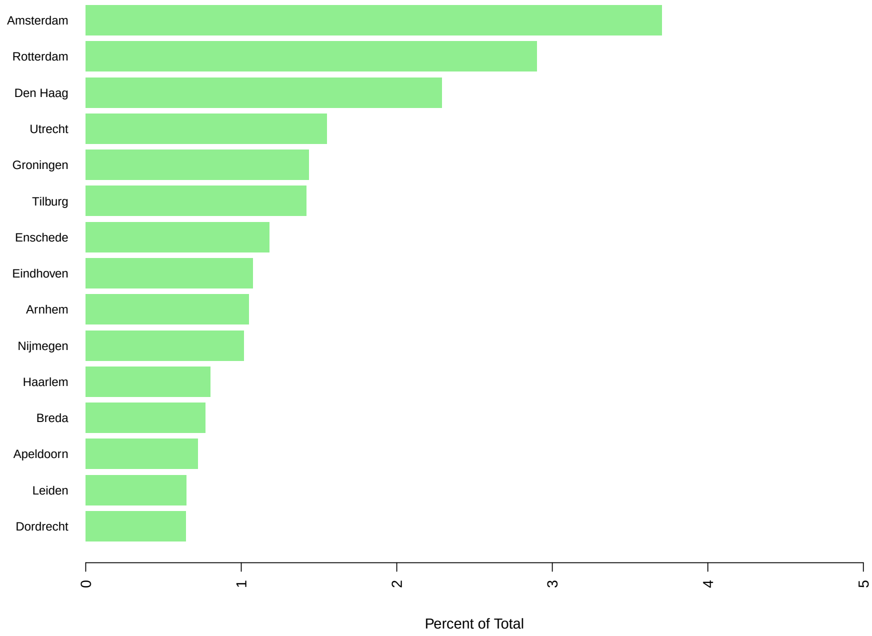


Figure 3-6: Top 15 cities with the highest number of Hyvers.

3-2 Groups and Topics

Each of the topics in Hyves contains many groups. Figure 3-7 shows the percentage of groups in each topic. We see that topics such as music, sports, brands have the largest number of groups. As illustrated, music has the highest number of groups in Hyves with nearly 200,000 groups followed by sports, brands, films, food and hangouts.

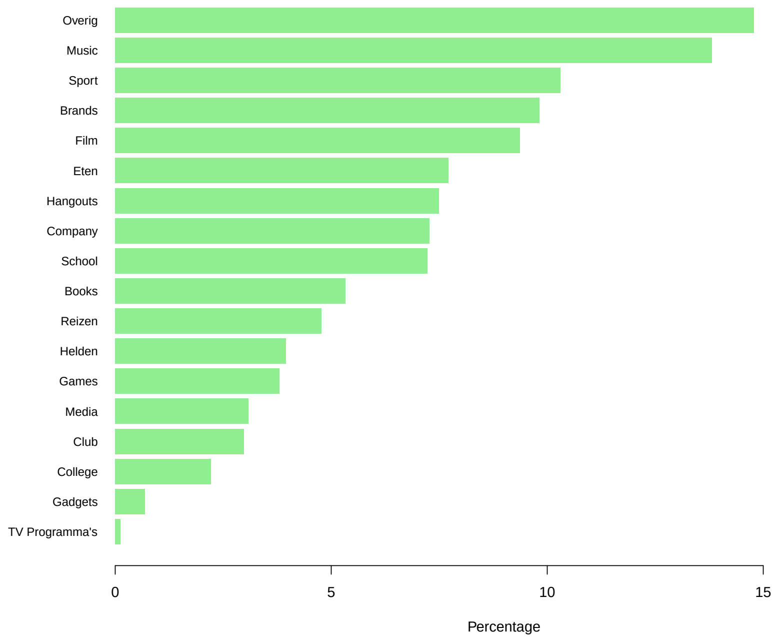


Figure 3-7: Percentage of groups in different topics in Hyves. Each bar shows what percentage of groups are in each topic.

The popularity of each topic is different. Figure 3-8 shows the percentage of users in each of the topics. Brands, music, food and films are among the most popular topics in Hyves.

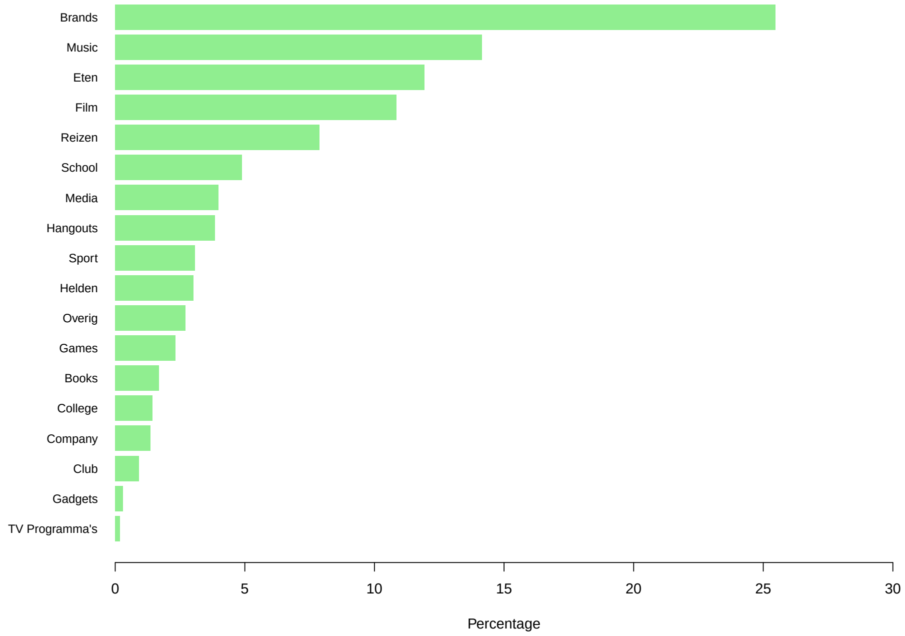


Figure 3-8: Popularity of each topic: each bar shows what percentage of users are interested in the topics.

3-3 Personal Information Disclosed in Hyves

In this section we focus more on the data available from Hyves user profiles and we examined the amount of information revealed by users that have a non-private profile. Figure 3-9 shows the percentage of information disclosed in Hyves for various attributes.

In our dataset, 70% of users reveal their birthday, 65% list their hometown, 45% relationship status, 41% school and 40% list their living status. Any of this attributes is likely to help identify a person and are more sensitive comparing to interests such as music and films. Ironically, such sensitive attribute are more available to strangers comparing to less sensitive attributes.

In case of relationship status, out of 1,537,528 profiles that have provided the attribute, 28% specifically selected the answer, “Don’t want to say it” showing that it is important for them not to reveal their relationship status on a social network rather than leaving it blank. That shows there is a concern for people in order to keep their relationship status hidden from public. More than 20% of users have selected “Unknown” as their relationship status. Among others, 20% are single, 15% are in a relationship, 10% are married, 6% are empty and the rest are other statuses such as open relationships.

We can also see the effect of age on relationship status. That helps us understand how con-

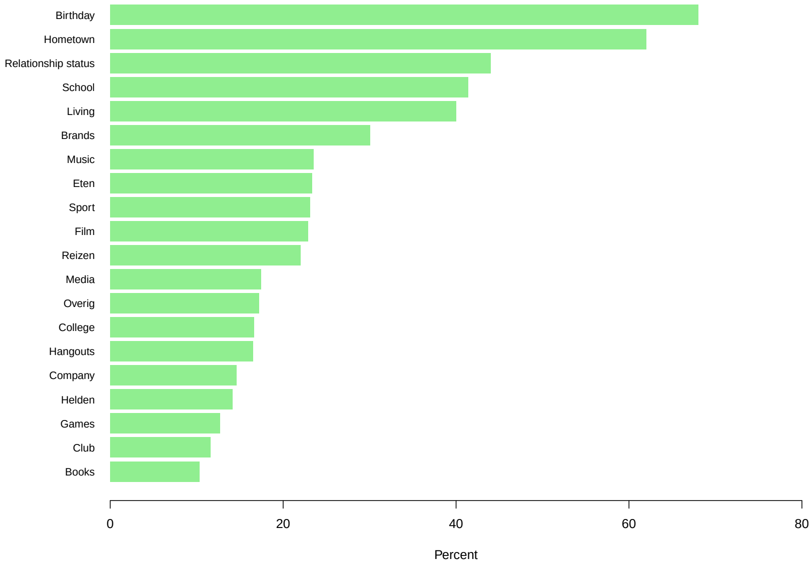


Figure 3-9: Percentage of Hyves users disclosing different types of information.

cerned are users about providing information about their private life. So, we consider profiles with each of the options single, married and in a relationship as public relationships while “Unknown” and “Don’t want to say it” as private relationship status. Figure 3-10 shows the fraction of users on their relationship status regarding the age. Each of the lines represents type of relationship status and each dot shows the fraction of profiles with that relationship in a certain age. Since the option “Don’t want to say it” does not exist among the options listed in Hyves anymore, we suspect that “Unknown” is used as a replacement. The plot also motivates this claim as both lines have very similar shapes as the age increases. Another observation is that until age 10, users that have provided their relationship status are much fewer than the other categories. Since it is unlikely to have a relationship status under 10, we suspect that profiles in this age have entered unreal value. From the age 10, we can see that the number of profiles with a valid relationship status increases significantly and about age 15 it crosses the other lines. On age 20 the value stabilizes until about 60 where the red line drops again and after 80 it stays closer to ’Unknown’ line. We interpret this drop because of users who enter invalid age value since it is unlikely to have a high number of users over 80. The reason is that as the age increases, number of users decrease significantly but after the age 80 it increases strangely. Since it is a mandatory field for registration, a fraction of users have provided an invalid and high value for their age.

The living status, refers to whom the user is living with. Out of 649,888 users who revealed their living status, 42% of them live with their parents which indicates that the majority of users are among young part of society and they are not independent. About 20% live with their partners, 14% are alone and the rest live with other family members.

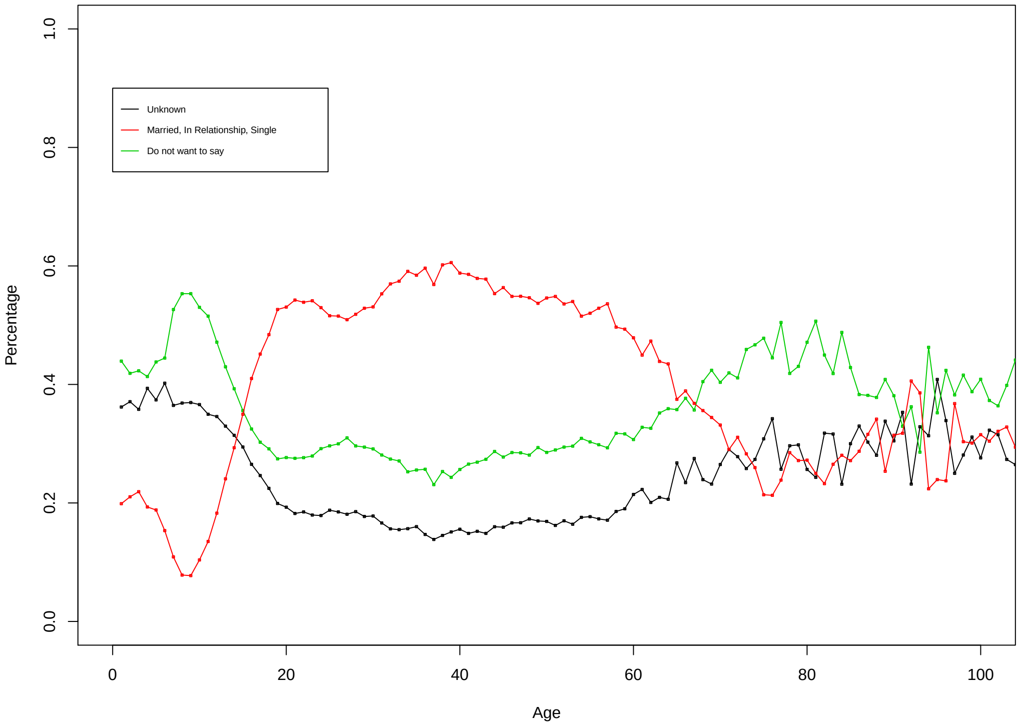


Figure 3-10: Relation of age and relationship status.

3-3-1 Profile Visibility

We were able to measure how many users have changed their privacy settings. To do so, we inspected the profiles in logged in state and logged out state and compared these states with each other. What can be inferred from the dataset is to know what amount of people have changed their profile settings to maintain a higher level of privacy. It is not feasible for us to distinguish about people that have chosen the options nobody, friends or friends of friends. So, we consider these three options as private.

Out of all the profiles, 41% of them have a profile that is visible to everyone on the web. This means they have not bothered tweaking privacy settings available in Hyves. Their profile page is accessible to search engines and web crawlers and can be easily collected. The reason can be either they do not care about having their information public or they are unaware of the features of the platform. Viewing the rest of profiles as an unregistered visitor is not possible because the profile is private and not viewable. This indicates that the privacy settings are configured and implies that for 59% of profiles, the privacy settings are a concern and have been changed. However, that does not necessarily mean the privacy settings are properly configured.

In order to see if the profiles are configured properly, we re-examine the visibility of private

Table 3-3: Visibility of user profiles based on privacy settings. Here “Others” consists of Friends of Friends, Only Friends and Nobody.

Category	Percentage
Everyone	41
Hyvers	12
Nobody, Friends, Friends of friends	47

profiles from the perspective of a logged in member (Hyver). If all of them are still private, that means by changing the privacy settings, a user has limited the access of profile to already known people (nobody, friends, friends of friends). If the profile is visible, that means the privacy settings is configured to limit the access of unregistered visitors only, meaning that any Hyver can view the profile. The possible interpretation of this action is to hide the profile from public listing and preventing search engines and web crawlers from collecting the information which is not fully effective and can be bypassed using different methods. The easiest way to do so, is to create a temporary profile and visit their profile as a Hyver. About 12% of the profiles are users with this situation and are visible to Hyvers. Finally, 47% of the profiles are protected from either unrelated Hyvers and unregistered users meaning that they have selected on of the restricted options (nobody, friends, friends of friends). This implies that at least about half of the users are concerned about their personal information and are aware of the means provided by OSN to protect their information. Table 3-3 shows the summary of profile visibility.

One of the misconception in OSNs is the term“friends of friends”. When someone thinks about friend of a friend, the real life equivalent comes to mind: a person who was with your friend in a bar. College friends of your girlfriend. They are nice people right? They are one step away so why not share information with them? The key question is how many people the average person can “reach” through friends of his friends. In average, a Hyves users has 218 friends. Assuming that the average friends list is interconnected and 50% of a user’s friends friends is shared with their friends we can measure the average number of reachable users by $\frac{218*218}{2} = 23,762$. This is surprising and it is unlikely that someone claim to have a trust in such a high amount of people. Hampton et al. [24] showed that the reach at 2-degrees of separation is estimated to be as high as 7,821,772 people (for a Facebook user that had a very large friends list that was not very interconnected). Their anlysis shows that OSN providers exploit the term “friends of friends” by not giving clear notices regarding the huge amount of reachable two hops away and users believe a friend of friend is a close relation.

3-3-2 Friends Network

Each Hyves user has 218 friends on average. Compared to Facebook, the number is slightly larger where a person has 190 friends on average [25]. For our analysis we were able to get the friendship relations of 406,139 users which resulted in about 87,837,000 friendship links. In Hyves 62% of users have more than 100 friends where this value for Facebook is 50% [25]. Figure 3-11 shows the histogram for number of friends in Hyves.

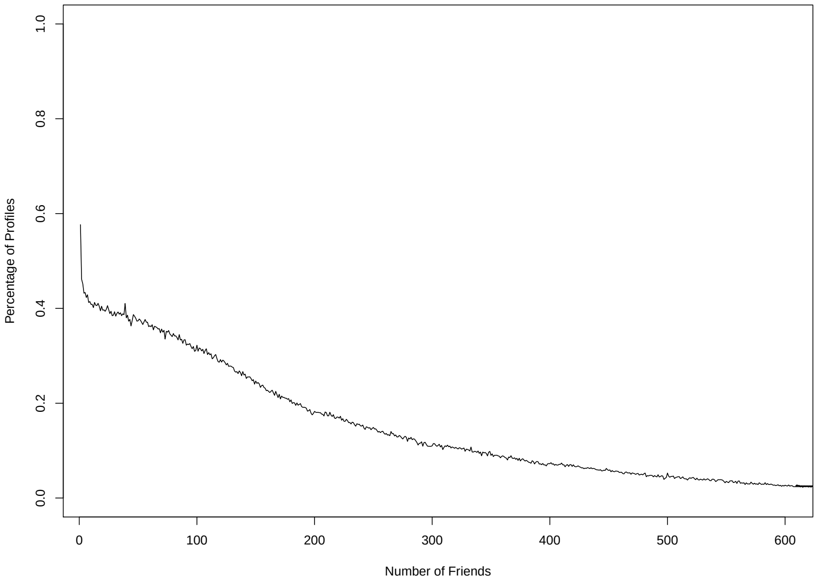


Figure 3-11: Histogram of the friend list size in Hyves. Users usually have large friend lists with an average of 218 friends.

The average number of friends is not the same for different age frames. Figure 3-12 shows the average number of friends with respect to age. For the range 0 to 5 and +90 there are strange peaks and as mentioned before indicates the users that are providing an invalid age. Since these two peaks are very similar to people in the range 11 to 20, we suspect that they are the teenagers that have entered a large or small invalid age. In chapter 5 we discuss these strange age ranges in more details. From age 11 until 20, average number of friends increases sharply and then starts to decrease until 70. This suggests younger people tend to spend more time on Hyves finding and connecting with their friends. For users with large number of friends, it is more likely to infer private information from their friends. That will also be discussed in chapter 5.

Since friendship relations are bidirectional in Hyves, by searching through the public profiles it is possible to find some friends of private users. In order to be able to compare how much it is possible to find the reverse friendship of a user, we use three metrics for each user:

1. Actual friends: Number of friends written in the profile page and indicates the total number of friends.
2. Available friends: Number of friends collected by our software.
3. Reverse friends: Number of friends collected by our software but in the reverse direction.

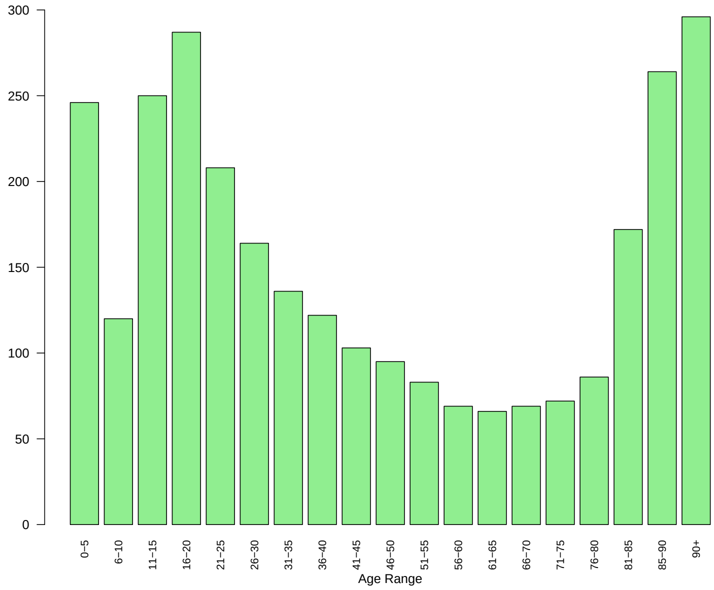


Figure 3-12: Average number of friends with respect to age.

Dividing Available and Reverse friends by Actual friends gives us the friends collection ratio and it shows the fraction of friends that were collected by our software. Figure 3-13 shows the histogram for friend collection ratio. The black line shows the Available friend collection ratio. That shows us for a high percentage of users, our crawler was able to fetch the whole connections. The red line is the Reverse friend collection ratio. The reason that there exist a ratio higher that 100 is that the process of collecting friendship relations and profile pages were not done at the same time. As a result, for some users the total number of friends in the profile may have changed leading to a ratio higher than 100. We can see that, although the Reverse friend collection ratio is smaller that the Available ratio, it is fairly possible to find the friends of any person by having the network graph. The Reverse collection ratio can be enhance and increased by crawling and adding more users in the friendship graph. It is observed that protecting the graph information should be mutual and it is not fully achievable by on person and on his own.

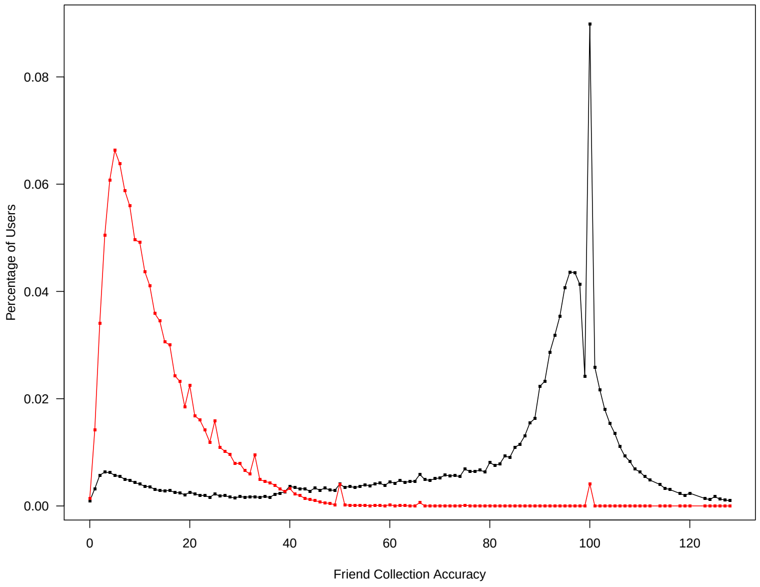


Figure 3-13: Friend collection ratio for Available and Reverse friends. The black line shows the Available friend collection ratio while the red line shows the reverse and is extracted by looking at friends. The reason for ratio higher than 100% is that the profile collection and friendship collection are done in different times so the number of friends has changed.

3-4 Summary

In this chapter we showed how easy it is to collect and analyze high volume of information from online social networks. We were able to catogorize users based on attributes such as age and hometown. It was also discussed that about half of the users configure their privacy settings and limit the visibility of their profile which implies there exist the privacy concern about giving out personal information. Yet, having a private profile does not guaranty that no one is able to find anything about you. That was simply shown by finding the the reverse friendship relations from the graph. In the next chapter, we go in more depth and see what private information can be inferred from public profiles in Hyves.

Chapter 4

Inferring Private Attributes

Social connections of a person define many of his properties. People are members of different communities in their life and each community describes an aspect of a person life. For example, for school students a majority of their connections are their classmates or colleagues at work form their own community. That suggests by knowing the people around a person, it is possible to guess some of the attributes of a person. In other words communities around a person define many of his characteristics.

Hyves allow users to select attributes such as school, company and interests such as music and brands. Also, by selecting these attributes users are grouped together so they can be in touch and communicate. While many users update their profile with such attributes, a fraction of them prefer not to have it visible to everyone but only to their friends. They configure their privacy settings and hide their profile. But what about their friends? Maybe their friends are not concerned about sharing their interests. As a result, a mixture of public and private profiles in the network is created. As we described 3-3-2 it is not a difficult task to find the public friends of a private profile since friendship links are bidirectional. By statistically analyzing the attributes of the friends it may be possible to infer the attributes of private profile.

Various methods for predicting user attributes can be used in OSNs. A very simple method to say something about private user attributes is to assign the most probable value of a specific attribute to all users that had not provided that attribute. For example, if the most probable age in the dataset is 20 then we assume that all those people that have not provided their age are 20 years old. This is a quite inaccurate because of heterogeneity of users inside a social network and provides very poor results. It does not make sense to assign a common attribute to all those who had not provided it.

For our analysis we use two methods. First we make predictions based on the friendship links inside the social networks. Then, we use *association rule learning* in order to discover interesting relations between user groups in the network. By generating association rules based on friendship relations, we then make predictions about user.

4-1 Predicting Attributes Based on Friendship Links

In this section we focus on two types of the attributes that users provide in Hyves: age and group memberships. The process is done by focusing on friends of a user and finding their similarity to him.

4-1-1 Age

An attribute that can be used for measuring the predictability is the age of users. As shown in section 3-1, a high number of Hyves users provide their age. To make a prediction about age and verify the results, we first see how the age differences for each of the friendship links are distributed. This can tell us in the first step how likely it is that people connected in the social graph are similar in their age. To do so, we measured the age difference of more than ten million friendship links in the graph and plotted the histogram of the absolute differences in Figure 4-1. As shown in the plot, 20% of the friendships in the graph have the same age and about 30% of links have only 1 year difference. This suggests that the majority of friendships is made between people that have equal or very similar age.

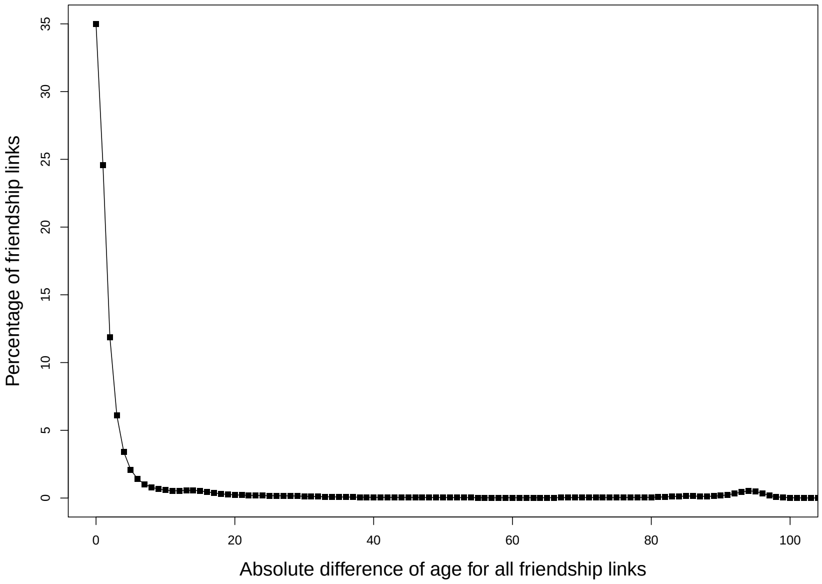


Figure 4-1: Histogram of age difference for all friendship links where the age of both sides are available. The total number is 11,444,390.

Since friendship links show close age differences, we make a prediction about the age of a person by picking the most porbable age among friends. We expect that users are more in touch with people of their own age. One of the questions arises is that whether this prediction is similar for different age groups or not. The answer can tell us whether we can make such predictions for all the users at a similar level of fidelity or if it depends on their age. Also, this is important as when predicting other user attributes and can tell us if this method can be applied to all the users or only a certain range of age. In chapter 5 we present the evaluation of this method.

4-1-2 Groups and Topics

Groups expose different characteristics of users in various topics. In this section we are interested to know how is it possible to make predictions of user groups based on different topics using their friendship relationships. Since we have collected the group memberships and the friendship links, for each user we can see what are shared with friends.

In contrast to age where friends tend to have high similarity to their friends user groups does not have such an overlap comparing to all friends. Users have broad range of groups and may have different preferences in their groups comparing to their friends. Figure 4-2 shows the probability that a user's friend lists the same groups as the user. The x-axis is the percentage of groups shared with friends and the y-axis is the probability in log-scale. We can observe that if all the friends are considered when finding the overlap taking the average will result in a small overlap (red points). That means there exists very few users that share similar groups with all of their friends. When searching the highest overlap between friends and users there exist much higher number of users which share a high percentage of groups with at least one of their friends. The plot can tell us that there exists certain friends that share similar groups to a user. This motivate us to use the friendship relations in order to predict group memberships for friends.

We examine some of the topics and show how different topics can be inferred from friends of a person. Our selected topics are school, company, brands, music. The reason for this selection is that they are different in their nature. Brands and music are topics with very popular groups. For example, one third of all users are a member of group Nike or Lady Gaga. That means if we choose a random user it is quite likely that he is a member of such popular group. In contrast, the school and company are more personal attributes and the chance that a random person is a member of a certain school is very small. For example in Hyves dataset the chance that a random person is a member of the most popular school is less than one percent. This may suggests that for topics such as brands and music, if a high percentage of friends are selecting a brand, it does not necessarily show that the person is a member of that group because the group is popular by itself. However for school and company, if a noteable amount of friends select a group it can indicate that the person is a member of that group as well. These suggestion will be evaluated in chapter 5.

In order to predict the groups in each topic, we use an approach similar to predicting the

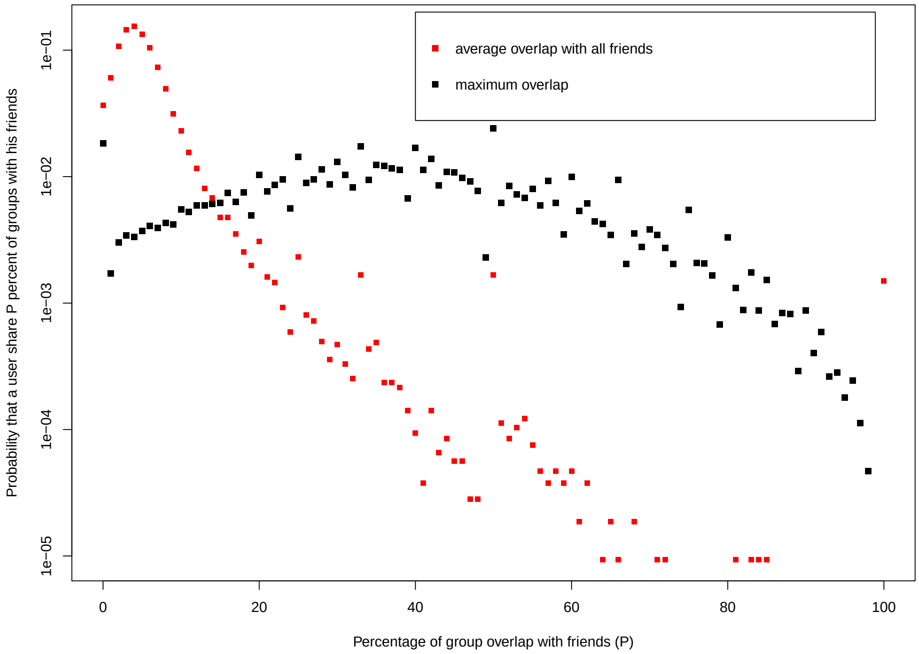


Figure 4-2: The percentage of similarity of a user's group to his friends group. The y-axis is log-scale. Red points show the average overlap between all friends of a users. Black is the maximal overlap with at least one of the friends.

age. For each user we predict a group in each topic based on the preference of his friends. By iterating over all friends we find which group has the highest probability among friends and select it as the predicted group for that user. Since users may have multiple groups in each topic, we limit the prediction to one group in each topic.

In the above method for each attribute, a certain fraction of friends share the most common attribute. For example, a user with 300 friends may have at most 10% of his friends having a common school hence the predicted school is only shared among a small subset of friends. We suspect the higher percentage of friends sharing an attribute, the more chance for the user to have that attribute as well. Also depending on the topic, it may require at least a percentage of friends to have an attribute in order to make an accurate prediction.

4-2 Predicting Attributes Using Association Rules

Association rule learning is a data mining method for discovering relations between attributes in large databases and is introduced by Agrawal et al [26]. The initial use case of association rule learning was *market basket analysis* where retailers try to understand the purchase behavior of customers and to find regularities between products. The process is usually done on large scale transaction databases recorded in supermarkets or shopping centers. For example, the rule $\{Milk, Diapers\} \Rightarrow \{Beer\}$ indicates that customers that buy milk and dipers are likely to buy beer as well. This analysis can then be used as the basis of decisions about marketing strategies such as sales promotion, product placement, cross-sellings and discount plans. For instance, putting beer close to milk and diapers can increase the profit of the market. Today, association rule learning is used in other fields such as web usage mining, bio-informatics and intrusion detection.

Users in an online social network and ther group memberships can be represented similar to a market basket. For example, users interested in the soccer club "Ajax Amsterdam" are also interested in "Amsterdam Arena" with a certain confidence. However, if a user is interested in "Addidas" in addition to Ajax "Amsterdam", he is likely to be interested in "Amsterdam Arena" with a higher confidence. Association rules allow us to statistically find what groups are interesting to users based on the appearance of the groups among users. As a result it can help infer the attributes of users based on the rules extracted in the network.

Our approach is to apply the association rule learning on Hyves group memberships and discovering the relations between users within a social network. And based on that we can see the relations between users within the social network and can be a mean to infer information about users on a global and local scale.

4-2-1 Definition

The original definition of association rule learning is proposed by Agrawal et al [26]. The problem is defined as follows: Let $I = \{i_1, i_2, ..., i_n\}$ be a set of binary attributes called *items*. Let $D = \{t_1, t_2, ..., t_m\}$ be a *database* of transactions. Each transaction t has a unique ID and contains a set of items in I. An association rule is defined as $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \varnothing$. In a market, this association rule means that if items in X are bought, it is likely that items in Y are also bought. The itemsets X and Y are called *antecedent* and *consequent*, respectively. We have adapted this definition to the users within a social network. In a social network, the items I would be the set of groups in the and each transaction t is a user. Users in the social network tend to be the members of different groups which are the items of a transaction. As a result each transaction (or user) has a unique ID (username) and contains a set of items (groups) in I. In this model, an association rule $X \Rightarrow Y$ means users that are a member of group X are likely to be a member of group Y.

To illustrate the concept, we show this definition by an example from our Hyves database. Imagine the set of groups are $I = \{Nike, Addidas, Spaghetti, Football\}$ and there are a set of members in these groups shown in Table 4-1. A "1" value shows the membership of a user in a group and "0" shows the opposite. An example rule in this database is $\{Nike, Addidas\} \Rightarrow \{Football\}$ which means people that are interested in Nike and Addidas are also interested in Football.

Not all the rules that are extracted in association rule learning are interesting. To find the interesting rules, they must satisfy certain user constraints called *support* and $confidence$.

Table 4-1: Example database with 4 groups and 5 users.

User ID	Nike	Addidas	Spaghetti	Football
1	1	1	0	1
2	0	0	0	0
3	1	0	1	0
4	1	1	0	0
5	0	0	1	1

These two metrics are defined as follows:

- **Support:** The *support* *sup*(X) of an itemset X is defined as the ratio of users which have X in their groups. For example, the itemset $\{Nike, Addidas, Football\}$ has a support of $\frac{1}{5} = 0.2$ since it appears in 20% of all users (user 1). Support determines the proportion of users which contain the group set.

- **Confidence:** The *confidence* of a rule $conf(X \Rightarrow Y)$ is defined as $\frac{sup(X \cup Y)}{sup(X)}$ which means that the percentage of users having X are also interested in Y. For instance, in the sample database, $\{Nike, Addidas\} \Rightarrow \{Football\}$ has a confidence of $\frac{0.2}{0.4} = 0.5$ which means that out of the users that are interested in Nike and Addidas, 50% of them are also interested in Football.

Support and confidence are important parameters in selecting the rules in the database. Rules with low support value may be happening only by chance. Moreover, a low support might not be definitive since it is not reflecting enough information about the dataset because its occurring for a small portion of users. A high support however, might prune some rules that contain valuable information about the dataset.

Confidence shows the reliability of a rule. In a rule $X \Rightarrow Y$, a higher confidence shows that its more likely that users interested in X are more likely to be interested in group Y. A high confidence might also be misleading. For example, imagine the following example: We want to find the relation of people who are interested in groups Versace and Football. Table 4-2 shows a sample dataset regarding these two groups. The data in this table suggests the rule $\{Football\} \Rightarrow \{Versace\}$ exist and people that like the football are also interested in the brand Versace since the support is 65% and confidence is 75%. However, regardless of being a football fan the fraction of Versace lovers is 80% . This implies that being a football fan has actually a negative effect on interest for Versace and provides that the rule $\{Football\} \Rightarrow \{Versace\}$ is misleading. To circumvent the issue *lift* value is defined as mean to determine the quality of a rule and is defined as follows:

Lift: lift value for a rule, is defined as $lift(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)sup(Y)}$. Lift is interpreted as the importance of a rule and shows how much the antecedent and the consequent are expected to appear with each other. When the lift value is larger than 1 it means that the antecedent and the consequent are apearring together more than expected hence the antecedent has a positive effect on the appearance of consequent. On the contrary, a lift value smaller than 1 means that the antecedent and consequent appeared together less than expected and as a result, the antecedent has a negative effect on occurrence of the consequent. A lift value equal to zero means that the antecedent and the consequent are independent variables and they do not have any effect on the appearance of each other. In the above example, $lif(\{Football\} \Rightarrow \{Versace\}) = \frac{65\%}{80\%*80\%} = 0.01\%$. Since the value is smaller than one, it indicates that being a football fan has a negative effect on being interested in Versace.

Table 4-2: Interest of 1000 users with respect to Football and Versace.

	$Football$	$\overline{Football}$	Σ
$\overline{Versace}$	150	50	200
$Versace$	650	150	800
Σ	800	200	1000

There are several algorithms to find association rules in a database. A well known algorithm widely used to find rules is *Apriori* algorithm introduced by Agrawal [27]. Apriori is based on breadth-first search and Hash tree structure to find the candidate itemsets efficiently. Depending on the support and confidence value, the number of rules varies. These two metrics can largely help us remove uninteresting rules from the result set of Apriori. Rules that are generated by Apriori, are in the form of $X \Rightarrow Y$ where the antecedent X is non-empty set and can have more than one element while the consequent Y is a set with a single item. Using Apriori, when support and confidence value is small, a lot of rules are generated. Not all the rules are useful and in many cases different combinations of antecedents point to the same consequent. To decrease the complexity of rule processing in our context, we limit this research only to those rules that have a single group in their antecedent and larger rules are considered for future works. To extract the rules, we used an efficient implementation of Apriori by Borglet [28].

4-2-2 Analysis of Hyves Using Association Rule Learning

In the previous section we adapted the definition of association rules to users of social networks. The group membership in social networks is similar to a market basket used in association learning. Different users are interested in different groups and by analyzing these memberships similar to a market basket, we would be able to find interesting rules about people and use it as a mechanism to predict attributes of users in a social network. By mapping this data set to association rule learning data and analyzing the rules, it is possible to make prediction about users.

In this section, we first describe a method to use association rules and see if they can help us on predicting user attributes by introducing the notion of *predictability*. Then on the second step, we propose a method to actually predict user attributes based on their friendship relations.

4-2-3 Predictability of Users

Our first approach on using association rules for Hyves dataset is to assess how much possible it is to rely on association rules. Association rules extracted from an OSN tells us which groups are likely to appear together among the users of the network. A large number of rules may be extracted and for each user we need to filter the rules and find the relevant rules.

Having extracted rules using Apriori, each user is compared to the set of rules and is determined how much of his groups are actually extracted in the rules. To have a value of predictability for each user, we define the *predictability* value. Two methods are used to define the predictability value which we called them the *naive* and *average-confidence* approaches.

- **Naive:** The naive predictability select the relevant rules for each user and and based on those rules, shows how much of his groups appear in the rule set.
- **Average-confidence:** The average-confidence predictability combines the relevant rules for each user and provides a measure that considers the confidence as well.

Figure 4-3 visualizes this concept in an example. Imagine Bob, John, Mary and Chris are a subset of users from the dataset and Figure 4-3 shows their group membership. Groups a, b, c, d, f are a subset of groups and they appeared as the groups of these users.

The first step towards defining the predictability is to generate association rules for the dataset using Apriori. In order to generate the rules we need to specify a minimum confidence C and and minimum support S. Running Apriori on the dataset with these threshold will result a set of rules in the form $X \Rightarrow Y$. We limit the rule generation to single sized antecedent and we can show a rule as $x \rightarrow y$ where x and y represent groups in the dataset. The rules are then stored and can be used for the next part of the analysis. Suppose that the following rules are extracted from the dataset:

$$r_1 : a \rightarrow c \mid c_1 > C \ , \ s_1 > S$$

$$r_2 : a \rightarrow b \mid c_2 > C \ , \ s_2 > S$$

The result indicates that there are two rules r_1, r_2 with a support and confidence higher than thresholds S, C respectively. Having the rules extracted, we can proceed to define the naive and average-confidence predictability measures.

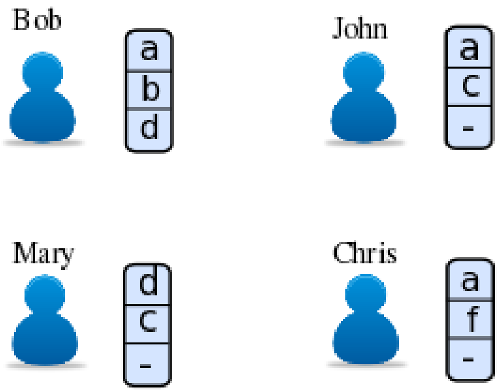


Figure 4-3: Sample set of users where Bob, John, Mary and Chris and their membership in groups a, b, c, d, f

Naive Predictability

The naive predictability value considers rules that are relevant to the user and then measures the percentage of user groups are predictable using that groups. Naive predictability value is calculated for each user using the following steps:

1. For each user u_i in the dataset find which rules *apply* to that user. A rule r_j applies to a user if both the antecedent and consequent of the rule are in the groups of that user.
2. Out of the selected rules, find all distinct groups that are occuring only in the consequent and call the set A_i.
3. Divide the length of A_i by the length of groups for the user G_i. Or simply:

$$v_i = \frac{|A_i|}{|G_i|}$$

   where v_i is the predictability value and indicates the percentage of groups for user u_i which are actually predictable.

To further explain the method described above, we go over each step. In the first step, only those rules are selected that have both antecedent and the consequent exist in user groups. The reason for such selection is that if the user is not appearing in the antecedent, that rule will not provide any information to help and find extra information about the groups even if the consequent is in his groups. On the other hand, if the the consequent does not exist in the user's groups, that means the rule is irrelevant for that user and Apriori did not considered this user during rule generation. Back to our example, r_1, r_2 apply to Bob, r_1 applies to John and no rule applies to Mary and Chris. When no rule applies to user i, v_i we consider the predictability as zero for that person.

As the second step, since both sides of the selected rules exist in the groups of user i, we are interested to know how many of the groups of the rule set actually appear in the consequent of rules. This indicates which of the user groups are actually predictable based on the ruleset that applies to him. For example, distinct consequent groups for Bob and John are $A_{Bob} = \{c, b\}$ and $A_{John} = \{c\}$ respectively. Notice, that group a is not appearing for any of them because there is no rule that makes a prediction about a.

Finally, by dividing the number of predictable groups by the total number of groups of a user, we can find the predictability of a user is based on the specified thresholds. Following our example, $v_{Bob} = \frac{|A_{Bob}|}{|G_{Bob}|} = \frac{1}{3} = 0.33$ and $v_{John} = \frac{|A_{John}|}{|G_{John}|} = \frac{1}{2} = 0.5$

One of the problems of the naive predictability value is that although the selected rules satisfy the minimum confidence, they are considered all equal in the sense of their predictability. For example, imagine that $C = 50\%$ and r_1, r_2 have the confidence of 90% and 50% respectively. Since in calculating naive value we are only counting the number of distinct groups in rule consequents, both r_1 and r_2 are counted as equal. However, the rules indicate that r_1 is much more likely to happen comparing to r_2 but this is not considered in calculating the predictability value. The average-confidence predictability value is defined to overcome this issue.

Average-confidence Predictability

The average-confidence predictability is a different approach comparing to the naive method. As described before the shortcomming of the naive approach is that all rules are considered as equal. However, different rules have different confidence values and it would make sense to consider that value in measuring the predictability. Since multiple rules can have similar consequents, for each group that appears in some rule consequents, we pick the rule that has the maximum confidence. If there are rules with similar consequent and confidence, the one with larger lift value is selected because a higher lift value indicates that the antecedent and the consequent are more related. Predictability is defined by taking the average confidence of the selected rules. The steps in finding average-confidence predictability is as follows:

1. Similar to naive approach, for each user u_i in the dataset find which rules apply to that user R_i. A rule r_j applies to a user if both the antecedent and consequent of the rule are in the groups of that user.
2. Out of the applied rules, for each distinct consequent, find the rule with maximum confidence. If there are multiple rules with similar consequent and confidence, select the one with higher lift value. Call the selected ruleset RS_i.
3. The average-confidence predictability is then computed as:

$$v_i = \frac{\sum\limits_{j=0}^{m} c_j}{|R_i|}$$

Where c_j is the confidence value for jth item in RS_i.

Back to our example, if the confidence of r_1 and r_2 are 90% and 60% respectively the average-confidence predictability for Bob and John would be 75% and 90% respectively and it would be zero for Mary and Chris because none of the rules apply to them. Table 4-3 shows the predictability values of the example for both naive and average-confidence methods.

Table 4-3: Average-confidence predictability value for the example data.

	Bob	John	Mary	Chris
Naive	33%	50%	0%	0%
Average-confidence	75%	90%	0%	0%

4-2-4 Predicting Attributes Using Association Rules

Although predictability value can tell us about the users and how well it is possible to extract information about them, it can not be used directly to make predictions about users. It can be computed only if we know the groups information of user and based on the known information predictability is measured. However, when predicting an attribute, we know nothing about the user being assessed. Here, the typical scenario of a prediction is to predict some unknown attributes of a given users by having the friendship information and the attributes of friends. After making the predictions it is possible to verify the result by comparing the the real value of the attribute and the prediction result.

Using association rules we define a method to predict groups of users of an OSN based on friendship relations. The main difference of this method with the method described in section 4-1-2 is that, the predictions does not pick only the most common group for prediction but a set of groups are predicted. For example, if a user has 10 groups in 3 topics the method of section 4-1-2 only predicts the most common group in each topic. However, a user may have multiple groups in different topics. Using association rules the predictions are made regardless of topic and a set of groups will be predicted.

In our method Apriori algorithm is run dynamically over each of the user's network and rules will be extracted specifically for each user. After extracting rules for each of the users, the set of groups that are appearing in the consequent of rules would be selected as the predicted groups. This method has several advantages: first, the computation complexity of rule generation decreases significantly because rules are generated only based on each person and his friends hence less amount of resources is needed. Secondly, it is possible to use lower support value in order to include smaller groups such as schools and companies. When running Apriori over all the groups because of the large number of rules it is not possible to lower the support from a certain value because of lack of memory and time. Third, since the rules satisfy a minimum confidence we are sure any predicted group is satisfying at least one rule with a certain confidence higher than the minimum confidence.

The minimum support plays an important role in our predictions because it determines the minimum percentage of friends needed to be present in groups for rule generation. This is helpful because different attributes may require different fraction of friends in order to make an accurate prediction. Also, the minimum support treshold allows us to filter out the groups with very small number of users to make statistically better prediction.

To verify the result of our prediction, the method is applied on the Hyves where their groups are available. After the predictions are made, we check what percentage of groups were predicted correctly using the rules. The accuracy of the prediction is defined as the percentage of a user's group which appeared in consequent of rules extracted from his friends meaning the number of groups that can be predicted from the association rules. For example, if a users has 5 groups and 3 of them exist in the consequent of generated rules the accuracy of prediction is 60% and that means 60% of groups can be predicted with a minimum confidence and support value.

Chapter 5

Evaluation

The previous chapter discussed the methods to predict user attributes in online social networks. The methods can be applied to any online social network where the friendship relations are available along with some of the attributes. We first used friendship relations to predict simple attributes such as age where a high number of users provided that in their profile. Then we described a similar approach for groups of users in different topics where the method can predict a single group in each of the topics. Then we showed that association rule learning can be adapted with group memberships in order to make prediction about a range of user groups of users regardless of the topic.

The evaluation of the proposed techniques is performed using Hyves.nl dataset described in chapter 3. The dataset contains the friendship relations for users where both the user and the friends have provided the assessed attribute to allow verifying the prediction results. The results of the analysis will be described in the following sections

5-1 Age

The first attribute to evaluate is age. As described in section 4-1-1 a high percentage of users in Hyves have provided their age and the age difference of friendship relations is small in the dataset. The age prediction mechanism was to pick the most common age among friends as the predicted age. In order to evaluate the method and see wether the prediction is accurate, we performed the prediction on those Hyves users with an available age on their profile. Then we measured the difference of the predicted age and the actual age and plotted a histogram of these differences in Figure 5-1. We can see that for about 40% of users the predicted age and the real age of users is exactly the same. When allowing up to ±1 year of difference the probability to predict the correct age of a user, by using the age of most friends, increases to 76%. This clearly indicates that users in a social network tend to be friends with people close

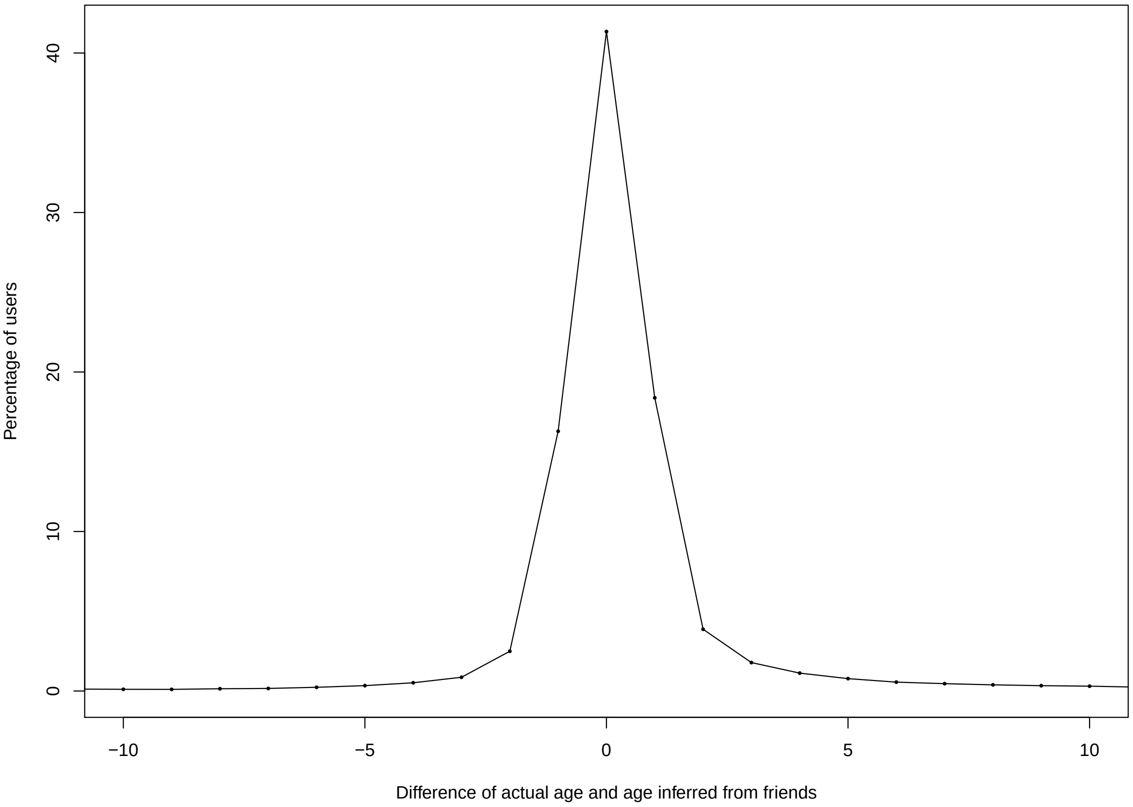


Figure 5-1: Histogram of difference of actual age and age inferred from friends. The plot considers the users which their friendship links and age is available (133,431 users).

to their age.

A question that raises is if this result holds for any person in the network or not. Usually, older people know more people in different age ranges where the teenagers mostly hang out with their school and college friends. To see the effect of age ranges, we have plotted the absolute difference of the real age to predicted age for various ranges in Figure 5-2 and Figure 5-3. The plot contains interesting information about different age ranges. For the age group 16 to 20 the prediction is accurate for more than 55% of the users. Allowing ±1 different increases that value to 95%. The same results closely holds for age group 11 to 15. As we expected, a very high percentage of teenagers are friend with other teenagers of their own age. Mostly, they know each other from school and college where they spend a considerable amount of time together. For the age groups between 21 to 40, the accuracy of prediction decreases slowly which means that users tend to be friends with people of different ages and that makes the prediction less accurate for higher age ranges.

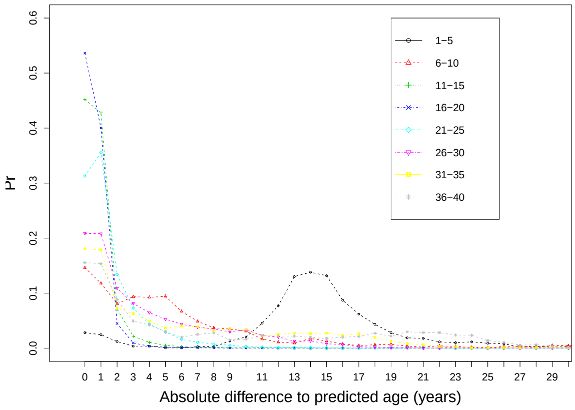


Figure 5-2: Absolute difference of real age and predicted age for age ranges below 40.

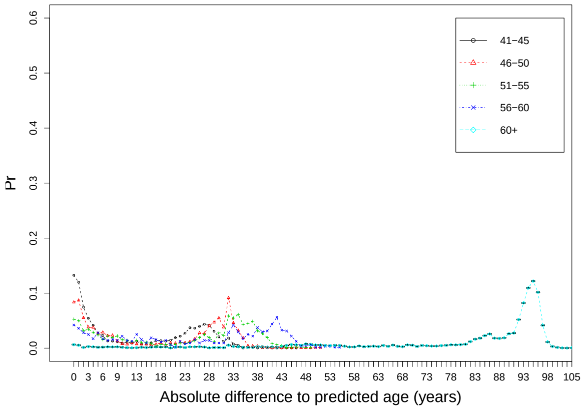


Figure 5-3: Absolute difference of real age and predicted age for age ranges above 40.

There are certain age groups which have peaks different to other groups. As an example, for about 78% of users in the age group 1 to 5, the predicted age has a difference of 10 to 20 years to the real age provided in their profile. It is obvious that a person with an age between 1 and 5 is unusual to have a profile in an OSN since he is not mature enough to use a social platform such as Hyves and we believe that users in this range are not providing their real age. In the plot, about 15% of users between 1 and 5 have a difference of 15 years with their predicted age. That means most of their friends are between 16 and 20. As we described in the previous paragraph, 95% of users in this age range can be predicted accurately. Consequently, we can claim that for 15% of users in age group 1 to 5 which have 15 years difference with their predicted age, it is still possible to predict their age with an accuracy of 95%. Using this approach we can claim that users which specified an age between 1 and 5 are almost 10 to 20 years away of their real age. Of course there is no way to verify this finding except asking the users and it is out of the focus of this thesis.

Another observation is that for age groups higher than 40, a new peak starts to appear at higher difference values. For example, for age group 46 to 50 about 8% of users are predicted accurately but for 10% of them the difference is 32 years which means 10% of their friends are between 14 to 18. Based on Figure 3-12 the average number of friends older than 40 and younger than 80 drops significantly and it is less than 100 friends while for other age groups this value can reach higher than 250. This is one of the reasons of poor prediction results for high age groups. They do not have too much friends and the small friends list can not provide enough information about their age.

The plot suggests that for users with age higher than 60 the prediction is not accurate at all because the difference of predicted age and real age is very large. However, there exists a peak at age 96 for this age range. Since the age can not be negative and the absolute difference is higher than 96 it means that person is at least 97 years old and most of the friends are one year old which is very unlikely. As we claimed before, users which are between 1 to 5 are most likely to be teenagers and as a result, we can claim that for users with age of 97 in their profile are mostly teenagers. Again there is no easy method to verify this calim.

5-2 Groups and Topics

Similar to age prediction, we predict the groups of users in each topic by selecting the most common group among friends. Four topics are used for predictions: school, company, brands and music. For each topic the the most common group is selected as the predicted group for the user. To verify whether the prediction is correct, the results is then compared to real groups of the user in his profile. We have compared the accuracy of our predictions based on age and percentage of friends having the similar attribute.

Figure 5-4 plots the accuracy of prediction for different age ranges and for different topics. Observations show that predicting groups for different age ranges is dependent on the topic. Typically for each of the topics, the age group 1 to 10 and +70 are not very consistent. As we described in previous sections, users in these age ranges are inconsistent and it is unlikely that their age is valid. As a result, the group prediction would not be very accurate and varies for these age groups.

We can observe that the prediction of brands can reach an accuracy up to 70% for age group 10 to 20 and then it decreases for higher ages. For school (red line) the predictions behavior is similar to brands but with a lower accuracy. The lines for brands and school suggests that the prediction becomes less accurate for older people. For schools we believe that older users tend less to update their school long after finishing that and since many of the friends have not updated their school in their profile, the accuracy is small. For brands, we suspect that younger people are more attracted to different brands rather that older people.

For company, we can see that the accuracy of predictions is less than 40% for users younger than 20 but it increases after the age 20 because most people get a job after finishing school and college.

The predictions accuracy for music is consistently small (about 30%) until age 20. After that it increases as the age increases and reaches even 60% for age 65. That indicates music taste as broader for younger people comparing to older users.

Our method of prediction is not accurate all the time especially when the fraction of friends with a common attribute is small. Also, it is important to know what percentage of friends should share a similar attribute in order to make a valid prediction. 5-5 shows the percentage of friends having the most common attribute against the accuracy of prediction. The first observation is that the accuracy of predictions highly depends on the percentage of friends having the predicted group. As shown in the plot, for all the topics if more than 50% of the friends have the predicted group the accuracy is higher than 50%. We can see as the

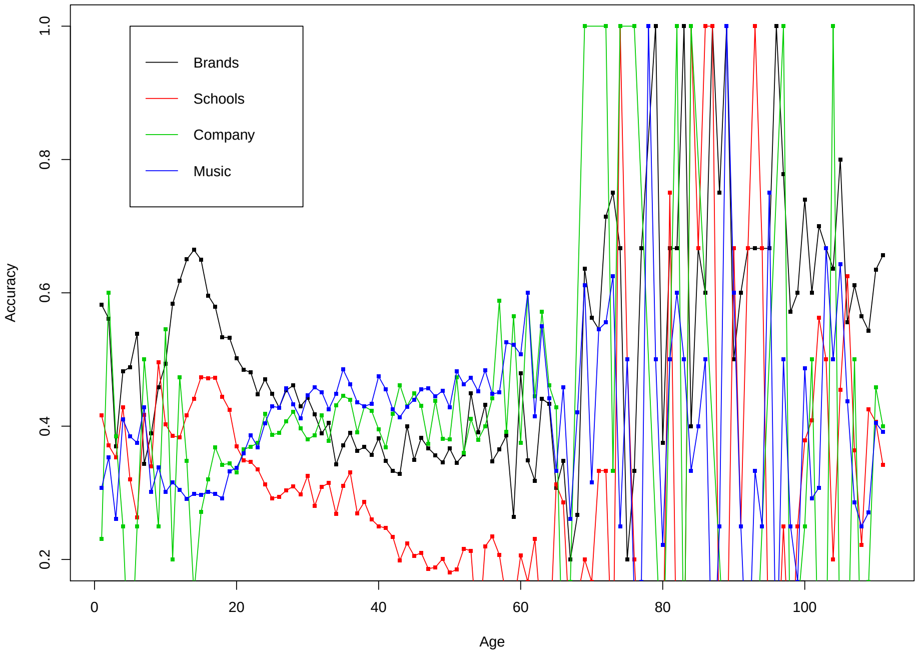


Figure 5-4: Accuracy of groups predictions based on age of users.

percentage of friends increases, the prediction becomes more accurate and for some topics it goes higher than 80%.

The plot shows that prediction accuracy varies for different topics. Music and brands are very similar to each other and the accuracy is smaller comparing to school and company. One explanation is that topics such as music and brands are more general topics where the groups are very popular among users. For example, a third of users are interested in brand Nike. However, if a small fraction of friends are interested in Nike, that does not necessarily mean the person is also interested in it. For brands and music, the the prediction is accurate only if a high percentage of friends are having that group.

For school and company, the prediction is higher than brands and music. This is maily because school and company are more user specific topics. A predicted school may have very smaller number of members comparing to a predicted brand. So even if a small percentage of friends are members of the school, it is less likely to happen randomly and can indicate that the user has a relation with that group. This is clearly visible in the plot as we see when a school is predicted from only 20% of friends, the accuracy of prediction is higher than 60%. When percentage of friends increases more than 50% the accuracy is up to 80%. The same holds for company and the accuracy is even higher for company. That says, even if a small fraction of friends are in a specific company, it is quite likely that the person is in that company as well.

The group predictions shows that for certain attributes, even a small fraction of friends can be used effectively to predict the attributes of a private users. This results imply that some-

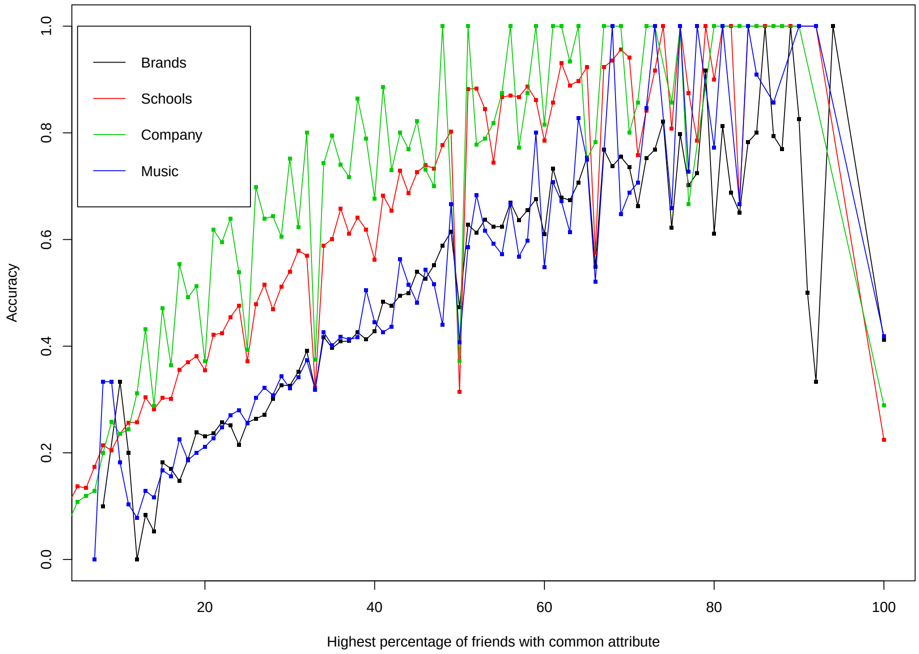


Figure 5-5: Percentage of friends sharing the predicted attribute versus the accuracy of prediction.

times, maintaining privacy is out of the control of a user because the friendship relations are already exposing information about the user.

5-3 Predictability of Users

We discussed association rule learning as a way to analyze the interests of users in online social networks. The naive and average-confidence predictability values are calculated based on the rules generated from the whole dataset. In order to extract rules we ran Apriori with different configurations on the dataset. The dataset contains 1,115,562 users from Hyves with their group memberships. In order to cover both middle sized groups as well as large groups, we have selected a minimum support value of $S = 0.005\%$ meaning that the for each rule should be satisfied among about 5.5K users. Lower support vale generates a very large number of rules which requires high amount of memory and resource to generate. To see how the number of rules change, we used three different values for confidence and extracted the rules under these thresholds. Table 5-1 shows number of rules for each of the configurations. Obviously, when minimum confidence increases, number of rules shrinks because not all the groups are occuring with each other so often.

Applying the naive predictability algorithm on the extracted rules provides predictability for each user. By running the algorithm on Hyves dataset, we observed that naive predictability

Table 5-1: Configuration of Apriori for global inference method. Minimum support for running the algorithm for dataset is $S = 0.005\%$

Confidence	50%	60%	70%
Number of rules	12231	4723	782

value is zero for at least 50% of the users based on different confidence values. The main reason is that for many users very small number of rules applies to them because the minimum support pruns a set of rules that may apply to users. As we shown in the example, if none of the association rules which satisfy the thresholds apply to a user, we can not say anything about that user and the predictability value would be zero. Table 5-2 shows what what percentage of users have a zero predictability under each confidence value.

Table 5-2: Percentage of users with zero global naive predictability.

Confidence	50%	60%	70%
Zero predictable users	50%	53%	60%

The rest of users have non-zero naive predictability. We can see that for a confidence of 50%, about 30% of all users (including those with zero values) have a naive predictability higher than 20%. When confidence decreases, naive predictability of users also decreases since smaller number of rules are generated. Figure 5-6 illustrates the histogram of naive predictability for different confidence values. We can observe that for smaller confidence values, a large number of users have high naive predictability value. The results imply that as the number of rules increases, predictability is also higher. As mentioned earlier, a problem of this approach is that smaller confidence values causes higher predictability value since there are more rules that match to user.

Proceeding to average-confidence predictability, we observed that 46% of users have a predictability of zero which is smaller comparing to naive predictability. Similar to naive approach we plot the result only for non-zero predictability values. Figure 5-7 shows the result for Hyves dataset. We observe that 26% of non-zero predictable users have a predictability higher than 20%. There is a peak on the graph for values between 50% and 55%. The reason for such a peak is because we are running Apriori with a confidence value of 50% and since there exist a fraction of users that have small number of rules which only satisfy the minimum confidence, taking the average confidence for rules with minimum of 50% causes the peak.

The predictability metric shows that although for a set of users the value is high, however for a large fraction of users the value is zero and indicates that the rules are not providing enough information about them. This is mainly because of the high support value of Apriori due to lack of processing resources which causes a lot of rules to be prunned. A lower support value will need large amount of memory to generate rules which is not available and our server was running out of memory. Also, a lower support creates a high number of rules. Since these rules are generated based on all the users and their interests, they cannot be used as a mean

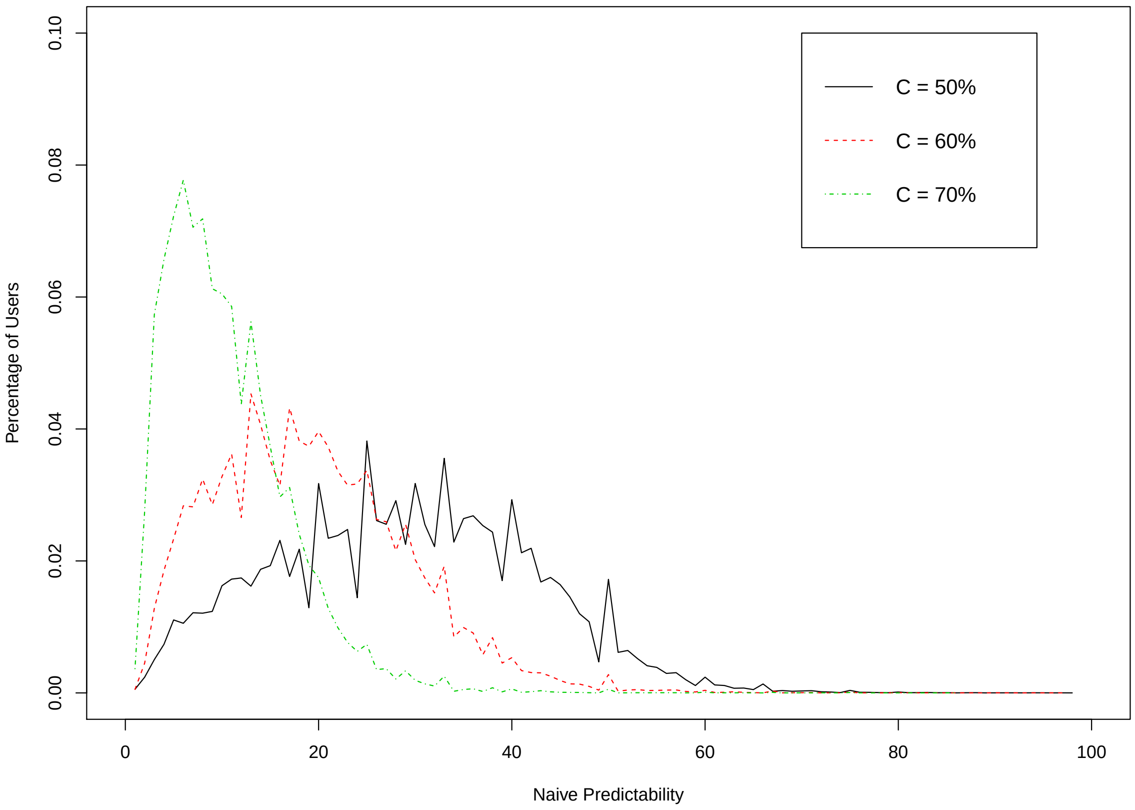


Figure 5-6: Naive predictability. Almost half of of the users have a zero predictability value (due to high support) which are not shown in this graph. When confidence increases it is less likely to predict about the users.

for predicting a specific user attribute. The reason is that given a private profile and the globally extracted rules, there is no way to select a set of rules for predictions since a lot of rules are extracted from the dataset and there is no way to distinguish them between the users.

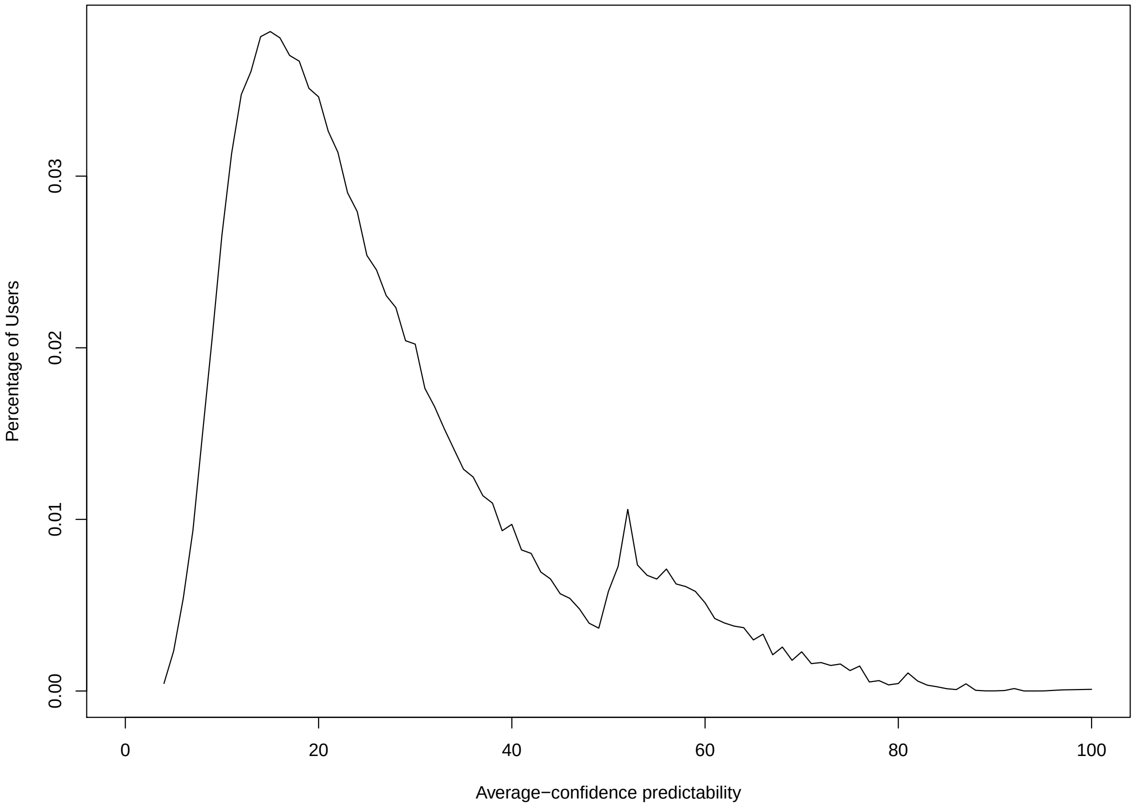


Figure 5-7: Average-confidence predictability for global inference.

5-4 Predicting Attributes Using Association Rules

In the previous section, all the rules were first generated using Apriori and then predictability value was computed based on the pre-generated rules. The rules are then uniformly used to find predictability for each of the users. In this section we present the results extracted from the method described in section 4-2-4. The process is done over the set of users which their friendship links are available.

As explained in section 4-2-4 the accuracy of predictions is defined as the percentage of user groups that are predicted correctly. The minimum support should be selected carefully when running the Apriori on friends. On one hand, a small support means that the rules are generated based on the membership of small number of friends in groups. As a result the predictions might not be be very accurate. On the other hand, a large support may prun the smaller but more predictive groups. In section 5-2 we showed that when even if a small percent of friends share a group, it is likely that the person is also a member of that group. Based on this results we have selected the minimum support value $S = 10\%$ for the lowest support value.

Since the average number of friends is 218 in Hyves, a support of 10% means that on average more than 20 friends should be the members of a groups in order generate a rule using Apriori.

Figure 5-8 shows the histogram of perdiction accuracy for different support values. The x-axis represents the percentage of groups predicted correctly and the y-axis is the percentage of users having that value. The plot clearly shows that when the support is 40% not a single group can be predicted for more than 25% of users and for the rest of users only a small percentage of groups can be predicted correctly. The reason is that it is less likely that 40% of friends share common groups because there exist different communities among friends. As the support decreases, the accuracy of our predictions increases. For a support value of 10% we were able to predict 50% of the groups for more than half of the users.

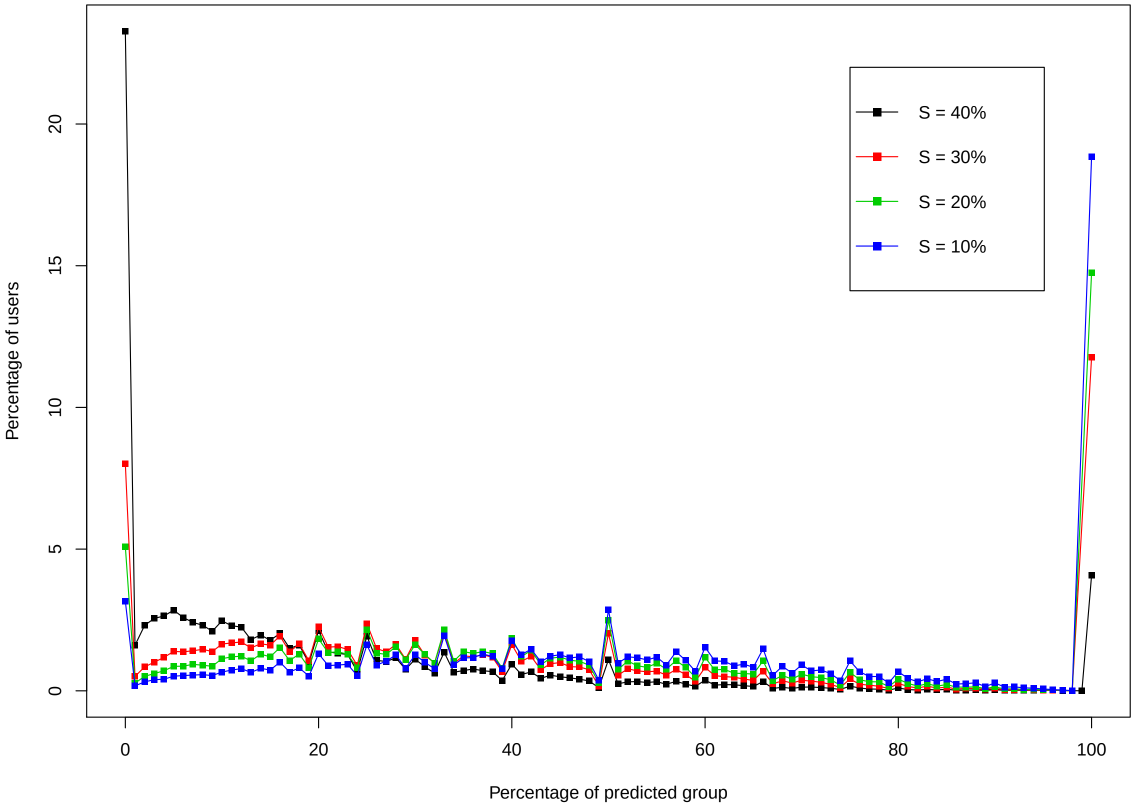


Figure 5-8: Histogram of prediction accuracy for different support values. As the support increases less number of users can be predicted because smaller groups are prunned.

For all the support values, there exist a high percentage of users which are predicted with a 100% accuracy. When the support decreases more users are predicted with 100% accuracy. For example for a support value of 10 about 20% of users have 100% accuracy. These users are mostly the ones which have smaller number of friends and cause the Apriori algorithm to generate the rules based on less number of friend. We have plotted the accuracy based on the number of friends in Figure 5-9. As we can see for all support values the accuracy is smaller for users with higher number of friends. There is a drop on the average number of friends when the accuracy is 100%. That indicates the predictions are made based on smaller subset of friends which share similar interests. For example, when the minimum support is 10% the average number of friends with 100% accuracy is about 50. That means when generating the association rules it is enough to have 5 friends in a group to predict that group. That explains why a high number of users have an accuracy of 100% because the rules are generated based on small subset of friends. When the support is 40% a larger number of friends are required to create the rules and that pruns the rules with small support. The results can still be justified because we have shown in section 5-2 even when small fraction of friends provide an attribute it is possible to make good prediction about the person.

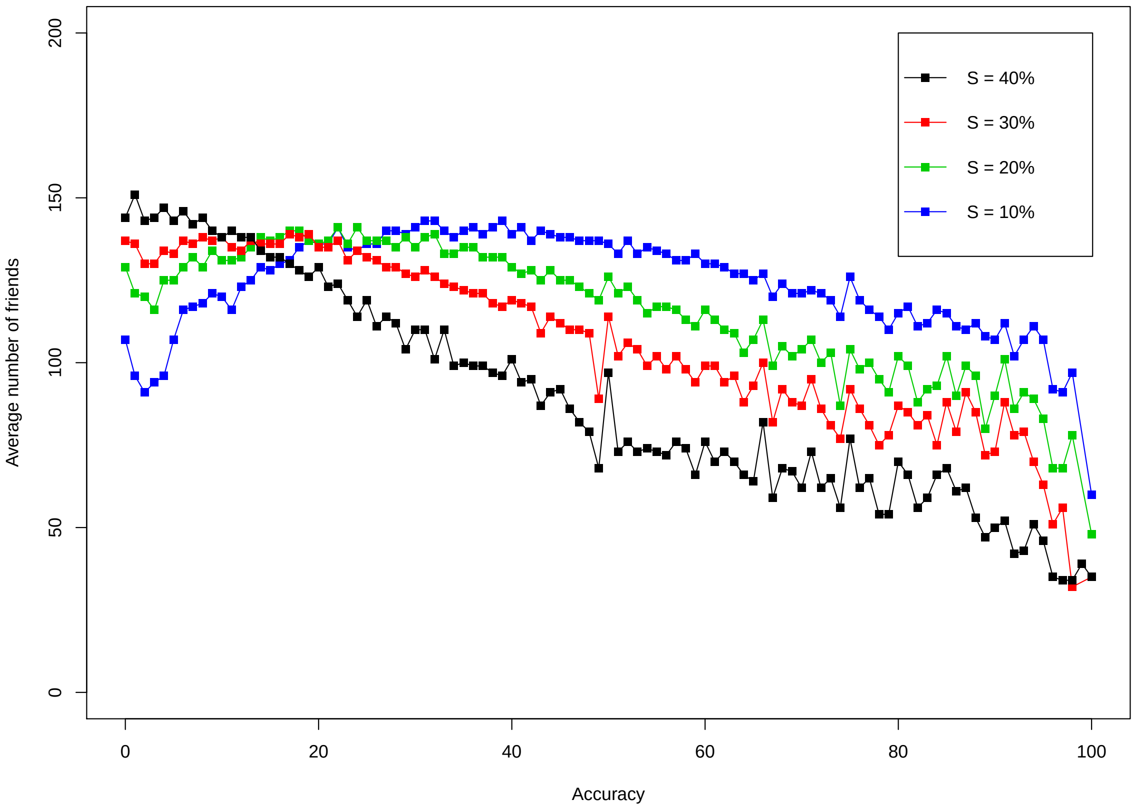


Figure 5-9: Accuracy of predictions versus average number of friends for different support values. It is observed that for users less number of friends predictions are more accurate.

Another question that my arise for our prediction mechanism is the *false accuracy*. Using association rules we predict a set of groups for a user and to verify the result we measure what percentage of user's group exist in the predicted groups. The problem is that sometimes the method predicts a large set of groups for a user but the person have only a small number of groups. If all the user's groups exist in the predicted set then the accuracy of prediction would be 100%. However, this prediction is not accurate enough because the number of groups that are falsely predicted are much higher than the the predicted groups. We define the false accuracy as the groups which are predicted falsely. For our result we observed that the false accuracy for any of the support values is less than 15% meaning that the predictions have small false accuracy. This tells us that the predicted groups are closely similar to the actual groups provided in user profile.

5-5 Summary and Discussion

In this chapter we evaluated the methods proposed in chapter 4 to make predictions about users. In section 5-1 we observed that it is possible to predict the age of users with an accuracy of more than 76%. However, the prediction accuracy varies for different age ranges. For younger people the predictions were more accurate comparing to older people because they younger people interact more with people withing their age. For higher age values users are connected to people with more various age ranges.

In section 5-2 we showed that the friendship relations can be used to predict other attributes such as school, company, favorite brands and music. The accuracy for the predictions were smaller than age predictions because such attributes are so diverse and the users' friends are among different communities. Still, it was observed that if certain amount of friends share an attribute it is possible to predict that attribute with a high accuracy. The accuracy also varies for different topics. For attributes such as company and school which are more personal even if a small fraction of friends have the same value it is more likely that the person have a similar value for that attribute. This was not the case for brands and music although they are usually popular groups with thousands of users.

Proceeding to association rule learning we defined the predictability metric in order to see the feasibility to infer private attributes of users. Predictability showed us what percentage of user's groups are exist in the generated rules. We used Apriori algorithm to generate association rules for all the groups of with certain confidence and support values. Due to memory constraints it was not possible to generate rules for sufficiently small support and as a result a high number of groups valueble for measuring predictability were prunned.

In order to predict the groups of users using association rules our proposed method was to generate the rules only among the friends of users and then the predictions are made based on the generated rules. The method predicts a set of groups which appear in the consequent of the generated rules for each of the users. The results show that by selecting a good support value it is possible to predict the groups of a user. A problem in this method is that the support should be selected carefully in order to prevent predicting groups based on rules extracted from very small number of friends. This usually happens for users with small number of friends.

A major problem in all the methods described in this study is that usually not all the friends of a person are useful in predicting certain attributes. In section 5-2 we showed that the average overlap of groups of a user with all the friends is very small for most of the users. However, it was also observed that for most of the users there exist a friend with high group overlap. Since our proposed techniques take all the friends into consideration, the irrelevant friends would have a negative effect on predictions. For example, when predicting the school of a user which is 15 years old friends which are older than 20 are less likely to provide useful information for predicting the school. If we are able to prun the irrelevant nodes from friends then the prediction accuracy of the attributes may increase.

A naive solution to find the relevant friends of a users is using brute-force or simply to try the prediction mechanisms on every subset of friends. Then the subset of friends with the highset accuracy would be the optimal subset of friends for prediction. We performed this method on a few number of users in our dataset in order to predict the school. We randomly selected 100 users with known school and small number of friends (between 40 to 50) because finding all the subsets of friends has an exponential complexity and requires large amount of time and memory. Due to these constraints we used the subsets with size of maximum 10 friends. For each subset of friends we compared the school with the user's school and measured what percentage of the subset have the same school. The subsets with highest similarity were then selected.

The results show that out of 100 users, more than 55 of the users have a subset of 10 friends where at least 50% of them share their school with person and 25 of them have a subset of 10 friends with exactly the same school. Also the subsets are very similar in their age. For about 14 users no school could be predicted. For each of the users most subsets were providing very small similarity with the person's school. However, there exist few but valuable subset of friends for the majority of users that could be used to predict the school. A further research question is to use an efficient method to find communities among a user's friends in order predict the hidden attributes. Depending on the intended attribute the communities may differ and different mechanisms may be needed to make community detecion. Finding these communities is a challenge which can be done in future.

Chapter 6

Conclusions and Future Works

Online social networks has become one of the most important media today. A majority of the internet users are using a variation of these platforms and try to communicate and connect with each others. OSNs has simplified the communications and formed communities faster around the globe.

However, online social networks can function as a double sided sword in people's live. While lots of users enjoy using social networks for free and fully trust OSN providers and other users, they might not be aware of the issues they are facing by puting their personal information online. By entering too much personal information in such services, people expose their private life to public without considering further consequences which can put them on risk. OSN providers may sell the user information to advertisers and even governments. Moreover, indepenent third parties may attempt to collect personal information for different activities without the consent of user.

In this thesis, we analyzed Hyves.nl online social network. We were able to collect nearly three million users in Hyves where more than half of the profiles were public. The profiles consist of personal information such as age, hometown, friendship links, relationship status, registration date and interests catogorized in various topics. Based on the visibiliy of the user attributes in their profile, we were able to measure what percentage of users have changed their privacy settings. It turned out that nearly half of the users have tweaked their privacy settings despite the fact that not all of them protect their personal information from public either because they are not aware the information is public or they prefer to share it with public.

In chapter 4 we proposed methods in order to make predictions about the attributes of users. Our methods rely on friendship relations of user and suggest that users are friends with people who share similar attributes and interests. The predictions were made by picking the most common attribute among friends and assigning that value to the user. Our evaluations show that for age, the prediction accuracy was quite high reaching to 76%. It is clearly visible that users tend to be friend with those closer to their age. This results show that the accuracy of predictions differ for users in different age groups. In younger age groups the accuracy is higher and it decreases when age increases.

We also used association rule learning in order to make predictions about the groups of users. Association rule learning is a widely used method in market basket analysis to increase the sale productivity in large databases. By using the group memberships of users as an input to Apriori algoritm, association rules were generated which satisfy a minimum confidence and support. The generated rules show the probability of presense of a group with a subset of other groups. This can help us measure the probability of user's membership in a specific group if they are member of other groups. Simply, association rules allow us to find frequent subset of groups. In order to make predictions about a private user, we generated association rules based on group membership of his friends. The extracted rules suggest a set of groups as predicted groups for each users. Comparing the predicted rules with the real groups of a user shows that for certain support values it is possible to predict more than 50% of groups for more than half of the users.

Some of the result of this thesis is published in [**?**]. The study can also be extended in different aspects. First, it is possible to apply the method on other datasets such as Facebook. This can give us insights about how social behaviour changes across different OSNs and whether they are the similar or not. Secondly, predicting the attributes based on friendship links can be extended to multiple hops to find the effect of friends of friends on the attributes of a user. This is tricky because on further hops number of users increases exponentially and the diversity of users increase. Third, not all the friends of a person may be useful the be used in predicting certain attributes. A key question is to find communities among user friends which expose the most similarities with friends. Using a brute-force selection of friends we observed that there exists certain communities which are more similar to the user comparing to other friends. Discovering efficient algoritms to remove weak ties and find the closer friends of a person can highly improve the prediction accuracy. This is the main roadline for future works.

In general, the main focus of this study was to practically show the possibility of finding information that are meant to be private. Users of social networks should be concerned about the information they are putting in OSNs. Once a piece of information or image is online, if not impossible, it would be very difficult to remove it. Personal information can be easily retrieved and analyzed to extract the hidden attributes of private profiles. Exposing this information can be embarrassing to users. Our analysis shows that even if the OSN is trusted and provides means to increase the user privacy, users should not fully trust their privacy tools and settings as a way of concealing their information. Friends expose too much information about each other and a user's friend list can be used to find different types of personal information. To reduce the privacy risks in online social networks users should be more concerned about the information they are entering in the first place.

Bibliography

[1] E. STEEL, "Facebook in privacy breach," http://online.wsj.com/article/SB10001424052702304772804575558484075236968.html.

[2] R. Clarke, "Introduction to dataveillance and information privacy, and definitions of terms," *Roger Clarke's Dataveillance and Information Privacy Pages*, 1999.

[3] abbas, "Google's eric schmidt on privacy," http://blogoscoped.com/archive/2009-12-07-n83.html.

[4] E. McCallister, T. Grance, and K. Scarfone, "Guide to protecting the confidentiality of personally identifiable information (pii)," http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf.

[5] "Data-theft worm targets google's orkut," http://blog.spywareguide.com/2006/06/datatheft_worm_targets_googles_1.html.

[6] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, ser. WPES '05. New York, NY, USA: ACM, 2005, pp. 71–80. [Online]. Available: http://doi.acm.org/10.1145/1102199.1102214

[7] M. McKeon, "The evolution of privacy on facebook," http://mattmckeon.com/facebook-privacy/.

[8] J. Kirk, "Europe chastises facebook over default privacy settings," http://www.pcworld.com/businesscenter/article/196232/europe_chastises_facebook_over_default_privacy_settings.html.

[9] B. Krishnamurthy and C. E. Wills, "Characterizing privacy in online social networks," in *Proceedings of the first workshop on Online social networks*, ser. WOSN '08. New York, NY, USA: ACM, 2008, pp. 37–42. [Online]. Available: http://doi.acm.org/10.1145/1397735.1397744

[10] J. Nagy and P. Pecho, “Social networks security,” in *Emerging Security Information, Systems and Technologies, 2009. SECURWARE ’09. Third International Conference on*, june 2009, pp. 321 –325.

[11] “Sophos facebook id probe shows 41% of users happy to reveal all to potential identity thieves,” http://www.sophos.com/en-us/press-office/press-releases/2007/08/facebook.aspx, 2007.

[12] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, “All your contacts are belong to us: automated identity theft attacks on social networks,” in *Proceedings of the 18th international conference on World wide web*, ser. WWW ’09. New York, NY, USA: ACM, 2009, pp. 551–560. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526784

[13] “The samy worm,” http://namb.la/popular/.

[14] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, “You are who you know: inferring user profiles in online social networks,” in *Proceedings of the third ACM international conference on Web search and data mining*, ser. WSDM ’10. New York, NY, USA: ACM, 2010, pp. 251–260. [Online]. Available: http://doi.acm.org/10.1145/1718487.1718519

[15] E. Zheleva and L. Getoor, “To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles,” in *Proceedings of the 18th international conference on World wide web*, ser. WWW ’09. New York, NY, USA: ACM, 2009, pp. 531–540. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526781

[16] “Facebook facts and figures,” http://www.website-monitoring.com/blog/2010/03/17/facebook-facts-and-figures-history-statistics/.

[17] “Face.com,” http://www.face.com/.

[18] J. a. P. Pesce, D. L. Casas, G. Rauber, and V. Almeida, “Privacy attacks in social media using photo tagging networks: a case study with facebook,” in *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, ser. PSOSM ’12. New York, NY, USA: ACM, 2012, pp. 4:1–4:8. [Online]. Available: http://doi.acm.org/10.1145/2185354.2185358

[19] A. Narayanan and V. Shmatikov, “De-anonymizing social networks,” 2009. [Online]. Available: http://arxiv.org/abs/0903.3276

[20] S. Zhong, Z. Yang, and R. N. Wright, “Privacy-enhancing k-anonymization of customer data,” in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ser. PODS ’05. New York, NY, USA: ACM, 2005, pp. 139–147. [Online]. Available: http://doi.acm.org/10.1145/1065167.1065185

[21] L. Backstrom, C. Dwork, and J. Kleinberg, “Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography,” in *Proceedings of the 16th international conference on World Wide Web*. Banff, Alberta, Canada: ACM, 2007, pp. 181–190. [Online]. Available: http://portal.acm.org/citation.cfm?id=1242572.1242598

[22] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," *SIGCOMM Comput. Commun. Rev.*, vol. 40, pp. 112–117, January 2010. [Online]. Available: http://doi.acm.org/10.1145/1672308.1672328

[23] "Social network analysis report-geographic-demographic and traffic data revealed," http://www.ignitesocialmedia.com/social-media-stats/2011-social-network-analysis-report/.

[24] K. N. Hampton, L. S. Goulet, C. Marlow, and L. Rainie, "Why most Facebook users get more than they give," Feb. 2012. [Online]. Available: http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

[25] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *CoRR*, vol. abs/1111.4503, 2011.

[26] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.*, vol. 22, pp. 207–216, June 1993. [Online]. Available: http://doi.acm.org/10.1145/170036.170072

[27] R. Agarwal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. of the 20th VLDB Conference*, 1994, pp. 487–499.

[28] C. Borgelt, "Efficient implementations of apriori and eclat," in *Workshop of Frequent Item Set Mining Implementations*, 2003.