University of Amsterdam
System & Network Engineering

Research Project 1

Synergy of social networks defeats online privacy

Authors:
Eleonora Petridou
eleonora.petridou@os3.nl

Marek Kuczyński
marek.kuczynski@os3.nl

Abstract

The popularity of online social networks is steadily growing and the users are willing to participate in a variety of networks and reveal sensitive personal information.The synergy of social networks can lead to a detailed profile of users. In this paper the risks of social network user profiling are examined. Moreover, risk scenarios are provided concerning the negative effects of data mining social networks. At the end there is a proof of concept regarding the method of crawling social networks.

`www.socialsynergy.nl`

February 7, 2011

Acknowledgments

We would like to thank Marc Smeets, Michiel van Veen, Ralf Bardoel and Marcus Bakker for their unwavering support and valuable assistance throughout the research project. We would also like to thank KPMG for providing us the opportunity to work at their offices and get acquainted with their professional environment.

- Eleonora and Marek

Contents

1 Introduction

Online social networks have become more popular over the last few years, the number of users of networks like Facebook or Twitter is growing exponentially. Many papers have been written about the technical details of data mining social networks, but little is known about the undesirable consequences someone could suffer when personal data from various social networks is combined and exposed to a third party.

Privacy risks like phising, identity theft and closer government monitoring have become more feasible with large sets of personal data freely available on the Internet. So far there has been little research into the privacy statements and settings that apply to social networks, while social networks are granting themselves more liberties day by day. Some of the social networks are also expanding into new areas like the advertising business, subscriptions for profiles with enhanced capabilities and data mining for marketing relevant information.

In this report we research the dangers of social data mining. We will provide practical proof of concepts about whether it is easily applicable to execute user profiling on a large scale and how the data could be exploited commercially. We will also research the countermeasures that can be taken against possible violations and provide our view on future developments and issues.

1.1 Research question

Our main research question is as follows;

What are the privacy risks associated with social network user profiling?

We will answer this question by investigating the following sub research questions;

- What is online privacy and why is it important to protect it?
- What is the current exposure of social network users to third parties and what controversial information is in the privacy statements of social networks?
- What business models do social networks follow regarding their users' data and privacy and what are possible future developments of these business models?
- What are the attack vectors on the existing social networks and how can obtained information be combined into something dangerous or valuable?
- What countermeasures can be applied to the problems we uncover in the proof of concepts?

The attack vectors will be demonstrated by the design and implementation of several proof of concepts.

1.2 Scope

This research will be focused on the privacy violations that social networks can cause for users. A proof of concept will be created to show how the data can be aggregated, including a description of how it could be misused by a third party. In the end, countermeasures will be listed together with a conclusion.

The following social networks have been included in the research;

- **Endomondo** - unique because it offers a software solution for sports tracking

- **Facebook** - by far the largest network, contains a lot of personal data
- **Foursquare** - a geographically based social network
- **LinkedIn** - the largest business oriented network worldwide
- **Twitter** - a very open communication based messaging service

The research group decided to adjust the scope of the research to the social networks mentioned above because they provide the most diverse information. This way the final dataset gathered from the total of the social networks during the proof of concept will consist of a diverse and complete set of information. This practice can lead to the best results when building a social profile of someone's online habits.

The research will be focussed on aggregating data from publicly available sources and not on gathering information from hacked or stolen accounts.

1.3 Related research

This project depends on some of the previous researches regarding the demographics and crawling possibilities of social networks. A lot of papers have already published about these subjects already, but very few of them focus on the ethical, legal and practical side of such privacy violations.

For example, one paper [2] demonstrates how the public information of 1.2 million social network users was enumerated by the use of e-mail addresses. Another paper [3] describes how a simple crawl of a Twitter user account can be combined into a demographic graph of the users that follow a specific Twitter user account. One of the most intriguing papers [4] describes how the creation of zombie users on Facebook can be automated in order to enumerate profile information from unsuspecting users.

2 Online Privacy

The official definition of privacy is the ability of an individual or a group to control the availability of information that is exposed about them. The importance of this principle is different for every person and culture, but the general perception of privacy is mostly based on direct human contact that people are used to in their day to day lives. In those scenarios, people have a pretty good understanding of the risks an eavesdropper could pose to their privacy, but this perception is not applicable to the internet.

The importance of the Internet has increased rapidly over the last few years, but a lot of the new features and mechanisms of the Internet are hard to comprehend by an average user. What is even worse, is that a lot of users cannot see the implications that the Internet could cause to their privacy and daily life. Recent surveys show that a lot of users are not very concerned with their privacy on the Internet and this is unlikely to change in the near future [2].

The Internet is an open global network where the laws and regulations of individual countries are very hard to apply, especially when a webpage is operated from abroad. This means that a lot of the protective matters against privacy violations within a country simply cannot be applied to the owner of a webpage. One of the few alternatives that is left open is blocking access to the webpage which can be considered as censorship [1]

2.1 Evolution Of Social Networks

A lot of the current giants in social networking originate from the period between 2003 and 2008 when the social network development was gaining its momentum. The term associated with this development is "web 2.0" [30], which refers to the concept of users contributing to and interacting with online content, instead of only viewing it. The term quickly became a hype and a lot of small startups involving social networking and social interaction were created. Some more successfull than others.

Before "web 2.0", a lot of Internet webpages were based on unidirectional communication from a company to a client or user. The average end-user was pretty safe within this setup, except when information had to be provided to the company as well, i.e. a credit card in order to charge the user for a purchase. This kinf of information is really important to keep secret, because in this example it is directly bound to a person's financial state.

Personal data that is shared through social network sites is perceived differently, since it does not confront the user with any direct consequences or commonly known risks. However, the advantage of social networking is that people can share and express some sort of personal affection, even though they might be located miles away from each other. Social networks also enable people to get new friends or connections and to maintain a relationship with them.

Another advantage of the social networks is that a user can receive a quick and brief overview of the life of his friends, contacts or associates without having to contact them directly. In this sense, social media are very egocentric because they require the users to expose themselves to a selected group of connections or the whole world before any interaction can take place between the users. Whether this development is a good or bad thing will not be discussed in this paper, but it does appear to have an impact on the way people communicate with each other in their real and online life.

2.2 Social Pressure

The established social networks have shown an exponential growth since their founding. This is partially due to the fact that a person's friends or connections are using it. In this section, an overview is given of reasons why social networks maintain a loyal user base.

2.2.1 Peer Pressure

A prime example of peer pressure can be found within the Facebook social network. It offers users the ability to interact in multiple ways, one of which is event planning and organisation. Within some usergroups it is very common to announce an event or to invite people to it through Facebook. This means that it is hard for a user to stop using Facebook, since it can lead to a form of social exclusion.

2.2.2 Special Features

The social network Endomondo offers users a complete and easy way to use sports tracker. The direct benefit of this service to the user is that the solution is free, accessible through an easy to use webpage and that the results of someones workouts can be shared with others. The drawback of this service is that the data is shared with a third party and that someone's workout data may be accessible by everyone if the profile settings are set incorrectly.

2.2.3 Discounts

Foursquare is a webservice that provides users the possibility to check-in to interesting spots, like bars or cinemas. Some establishments offer discounts or special promotions to people that frequently check-in at a business, which of course makes it very attractive to a customer to participate. It does however drag the user into a obligation to share the location and time of presence, which can then be abused either by marketing companies or criminals (i.e. burglars).

2.2.4 Communication

The power of Twitter is that it is an accessible and easy to use service. Because of its large and widespread user base, it enables both individuals or companies to broadcast a message to their followers and people searching for a certain keyword. Whether this is a good or bad thing depends largely on the content that a user posts on his Twitter page, but it is not hard to imagine that private or personal information can be abused in more than one way.

3 Characteristics of online social networks

The impact of social networks to society, especially regarding the younger population has been very strong. There is a variety of online social networks depending on the region as well as their features. Social networks have many characteristics in common but each of the social networks covered in this report has a special feature that makes it unique. The networks also differ in their purpose as some of them concern personal connections while others have a business orientation.

It is not well known how social networks generate revenue, so in this section a brief rundown is given on the information the research team could acquire. While the proof of concept and some of the scenario's are focussing on third party involvement in social networks, this section attempts to clarify in what degree the second party is involved as well.

The research conducted is directed towards social networks with different functionality, diverse data and rather high popularity. Observations within this chapter are based upon the privacy statements of the websites from January 2011. The privacy observations are mostly aimed at statements that stand out when compared to regular privacy policies.

3.1 Endomondo - Sports Tracking

Endomondo is a sports oriented social network. The idea behind it is to bring people together that are involved in sports by offering them free GPS tracking software. Users can install the software package Endotracker on their GPS mobile phones or on a Garmin GPS watch. Using a personal profile they can set up their sport goals, track their sports data and monitor their personal statistics. The whole route that a user follows during his workout is available at his profile page in real-time through a Google Maps image on the Endomodno website. Apart from that, other information concerning the kind of sport he/she did, the duration of the workout and the calories burned are also available.

The users can add other users as friends and comment on each other's sport achievements. The users can challenge their friends to achieve a better performance and upload photos from their exercise routes. Moreover, Endomondo can be connected with a user's Facebook or Twitter profile and every update can be made automatically available also to these two social networks as well.

Endomondo Privacy Observations

Endomondo users should be aware that the default privacy settings for the account make every piece of information (except for weight) visible to everyone. That means the route someone follows during the workout is made public unless the users chooses to disable this.

Endomondo wishes to protect the stored personal data, so they have taken several security measures. However, the disclosure of information publically is done at the user's own risk. Endomondo cooperates also with Facebook and Twitter so that the workouts of the users can be shared onto these social networks. But in this case, the privacy policy of Endomondo is not valid any more.

Endomondo states that they have the right to anonymize the data the users puts on their profile and to copy, process, use, public display and distribute this information. It is not certain

for what purpose the data is used or in what way the data is anonymized. There is a possibility, despite the fact that the data is anonymized, that a marketing company for example could figure out the real name of the corresponding user.

It should be mentioned that regardless of the privacy options Endomondo offers, the kind of data itself that the users upload can create certain real life risks. The fact that the daily workout of a person is displayed publically by default could be exploited by criminals i.e. burglars.

3.2 Facebook - Social Contact Oriented

One of the most prominent social networks is Facebook counting over 500 million users [8]. Facebook allows its users to create a personalised profile which can contain photos, videos, personal information and educational information. Users can also express their interests as well as their political and religious views. Interaction takes place because every user can add other users as friends, exchange messages and see each other's updates. Moreover, Facebook gives the opportunity to join common interest groups. Facebook users can organise events and invite other users. The latter can make public if they are attending the event or not. Facebook also offers an instant messaging service [9]. A new feature of Facebook is Facebook Places, which is available now in USA, UK, Japan and very recently in Europe. Facebook Places allows the users to check in at the places they visit and also tag other friends. Furthermore, it offers the possibility to locate friends that are nearby and gain discounts that are available near the spot. Apparently Facebook was motivated to promote Places because of the success of Foursquare. Another feature of Facebook is Facebook Connect. The users can use their Facebook credentials to log in to other websites without creating a new account.

Facebook Privacy Observations

Facebook has a very extensive privacy policy of 5,830 words. There have been many controversial issues and criticism against Facebook and the privacy policies it published from time to time. It is important to mention how Facebook uses the content that users upload as well as how it shares that information. According to Facebook's privacy policy [10] the uploaded information is used for personalised advertising. This means that Facebook offers to the advertisers enough data so that they can determine their target group. Facebook has the right to provide information that the user chose not to be public in his profile, such as the year of birth.

Moreover, when a user clicks on an advertisement, a cookie may be stored on his/her browser so that the advertiser is aware that the specific user fits the target audience. Facebook also uses collected information from a user and the user's friends to make advertisements more suitable and to increase the viewers interest. This is succeeded by putting a users name or picture next to an advertisement and mentioning that this user likes this application or product. Users have the option to opt-out but the feature is by default opt-in.

It is also worth mentioning that Facebook gathers data from other websites in order to have an archive of how users respond to specific advertisements. They claim in their privacy policy that they anonymize the data after 180 days, but they do not mention how the data is used before and after this step. There is no information on how the data is anonymized either.

Moreover, Facebook Connect has been considered as potentially dangerous because it actually lets one company rule a great deal of the social web. It is heavily used because users find it easy to log in everywhere without having to remember a pile of credentials. However, Facebook Connect allows Facebook to track the behaviour of the users in numerous websites without them noticing that.

It is not only Facebook itself that might use the information in an undesirable way, but also the users. In Facebook's Statement of Rights and Responsibilities [11] is stated that no user has the right to collect information on other users using means such as scripts or zombie computers. However, there have been incidents of data mining even at the very first days of Facebook at 2005. Students from the Massachusetts Institute of Technology managed to enumerate 70.000 Facebook profiles [12]. Other universities have been doing the same over time [31].

Deactivating or deleting an account might be an issue as well. When someone deactivates an account, it is not entirely erased. This serves the users that may want to reactivate their account but for the rest of them that wish to stop using Facebook may be frustrating. The personal information is not public but is still there. Furthermore, if information has been shared with other users then this remains accessible.

Last but not least it is important to mention the policy that Facebook follows regarding the updates of their Privacy Policy. The users can be informed through Facebook Site Governance Page or if they log into their account but not by any other interactive method such as an e-mail. This means that inactive users will not be made aware of violations with their profile.

3.3 Foursquare - Location Oriented

Foursquare is a geographically oriented social network used on cell phones or other GPS enabled devices. Users can check in at different locations and see if their friends are in the same neighbourhood. They can also earn badges as achievements and become the "mayor" of a location if they have the largest number of checkins at this specific place.

In order to promote products or services, some businesses started with promotional campaigns involving Foursquare. As an example, some coffeehouses will offer a free coffee if a Foursquare user can prove to be the mayor of that venue. Foursquare users can also read suggestions and tips about what they could do in their direct proximity. Most of tips are added by the end users or the venue owners.

Foursquare Privacy Observations

Foursquare may use the personal information the users supply and the technical details from the GPS device to improve their service. They also reserve the right to use this information to customize advertisements, the e-mails a user may receive and the content that will be presented to each user within Foursquare. It is possible to opt-out from e-mail advertisements and personal information is never sold or rented to third party companies.

3.4 LinkedIn - Business Oriented

LinkedIn is a social network oriented at business related connections. It has over 90 million users worldwide of which approximately half is located in the United States. LinkedIn users can

set up a personal profile containing information similar to that of their job resume. Afterwards, they can connect to friends, colleagues or associates in order to build up their own professional network.

A standard LinkedIn profile may contain information about a person's professional life, including job history, education, certificates, group memberships and connections. Optionally, personal information can be added, i.e. phone numbers, Skype userhandle or e-mail address.

LinkedIn can be a useful resource for those that want to find a new job and for corporate recruiters that are looking for candidates to fill a job vacancy. There are three subscription options that can support recruiters with their work, i.e. by showing the full first and last name for third degree connections and by providing better search and filter functionality for new candidates. It also allows the recruiter to directly message someone beyond the reach of their professional network.

Job-seekers have paid subscription options to help them find a job quicker. They can i.e. contact a company recruiter directly or allow direct access for a recruiter to their profile information. Unfortunately, there are no statistics available for the amount of users using one of the subscriptions.

LinkedIn Privacy Observations

The LinkedIn privacy policy states that LinkedIn reserves the right to send promotional emails, there is no opt-out for this option. LinkedIn will not sell a user's *personally identifiable* information to third parties, but the statement does not explain how this data is anonymized or who could either receive it from them.

A user has the option to control the public exposure of his/her profile in search engines and on the publicly accessible parts of the website. There are three privacy levels a user can set for displaying contact information and group memberships, but there is no option to change the visibility of the job experience and positions a person has.

3.5 Twitter - Communication Oriented

The Twitter messaging service is based upon short text messages (so called "Tweets") which users can add onto their profile. The messages a user puts online are visible by everyone by default, but there is an opt-in option to reveal them only to the users that have a connection with the publisher (the so called "followers"). Twitter claims to have 175 million users as of September 2010 [13].

Twitter can be used either through the website, third party software or through a mobile phone application. There are many third party services that extend the sharing capabilities of

Twitter, like i.e. services that allow users to upload and share pictures with their followers. Developers can also call upon an API to send and retrieve information from Twitter, such as the followers of an account, possible GPS tagging locations and a real-time stream of tweets.

Third party developers have extended on the limited messaging capabilities of Twitter. One example of this is Twitpic. Twitpic enables users to upload pictures to the Twitpic page, after which a shortened link to the picture is automatically tweeted. Followers can then see the picture on the Twitpic page.

Twitter Privacy Observations

The Twitter privacy policy [14] starts by stating which parts of the information a user submitted during registration (like full name and user handle) will be visible from publicly accessible search results. The content posted on Twitter, including additional personal information like a biography or a profile picture, is shown publically by default. It is possible to disable public access to parts of this information though. Optionally, GPS coordinates can be sent to Twitter as well, but this option is opt-in. Twitter does not disclose any information to third parties, unless they consider it to be necessary for legal reasons.

3.6 Aggregated Information on Social Networks

In the following table we want to give an overview of the characteristics of social networks so that it is more convenient for the reader to compare them and gain a better understanding of their growth and valuation. We refer to attributes such as initiation year, amount of users, number of employees and revenue. It is not always possible to find information because these companies are not part of the stock market yet, so their financial statistics are not officially published.

	Facebook	**Endomondo**	**Twitter**	**LinkedIn**	**Foursquare**
Initiation year	2004	2007	2006	2003	2008
Revenue	$800m (2010)	no data	$150m (2010)	estimation $205m (2010)	no data
Profit	Yes	no data	no data	no data	no data
Employees	1700+ (2010)	15 (2010)	351 (2010)	no data	32 (2010)
Amount Of Users	500m+ (2010)	500k (2010)	175m+ (2010)	90m+ (2010)	5m (2010)
Minimum user age	13	0	13	0	13
Area served	Global	Global	Global	Global	Global
Type	Profile based	Profile based	Content based	Profile based	Profile based
Photos	Yes	Yes	No	Profile pict.	Yes
User content	Yes	Yes	Yes	Yes	Yes
Cost	Free	Free, paid options	Free	Free, paid options	Free
Banner advertising	Yes	Yes	Yes	Yes	Yes
User applications	Yes	No	No	No	No
Platforms	PC/Mobile	Mobile	PC/Mobile	PC	Mobile
Policy URL	Facebook	Endomondo	Twitter	LinkedIn	Foursquare

Table 1: Information about online social networks.

- *photos* - the user can upload
  → just a profile picture so that his acquaintances can recognise him
  → whole collections of photos to share his experiences

- *user content*
  → updating status
  → post messages, comments, thoughts

- *user applications* - Facebook
  → games and quizzes
  → negative aspect the access to personal information and friends

- *Policy URL*
  → links for detailed document of the privacy policies

3.7 Conclusion

The growth of online social networks is obvious when looking at the amount of users and the variety of the existing networks. Focusing the research on different kind of networks helps to gather variable data and compare various privacy policies. Endomondo, Facebook and LinkedIn claim to anonymize the data before they try to use them for advertising purposes. However there are unclear points in the privacy policy of the above networks that could allow them to sell the data the users upload to third parties without notifying the users. It has not been proved that they actually sell data but according to the boundaries set in the privacy policies, they could. Furthermore, the majority of the profile features is publicly exposed by default. The social networks leave to the user the responsibility to scan through the privacy settings and protect the data he uploads.

Independently of the privacy policies it should be mentioned that the kind of data some social networks request, such as the GPS coordinates and home address, are potentially dangerous even if they are not sold to third parties.

Another issue are the Facebook user applications. Whenever a user wishes to use an application he/she has to agree first that the application can access personal information of the users profile. Quite recently, Facebook announced that the applications can also access the mobile phone and the address of the users if permission is given [5]. After protests of the users Facebook withdrew this function temporarily. However, this shows that more and more data are exposed uncontrolled with the consent of the social networks.

4 Social Networks Business Models

The value of social networks keeps rising and it is important to see whether online social networks can generate money. Numerous investors have shown faith in them, one of the latest investment was made by Goldman Sachs involving $450 million[15] on Facebook. This means that investors do expect to get their money back. However, it is not only the investors money that maintain the social networks alive. In this chapter it is discussed how the social networks themselves have initiated revenue generating products and services so that they can become really profitable businesses.

4.1 Twitter Business Model

Twitter started a new advertising program in April 2010 that places contextually relevant advertisements to the user's search results. The program is called *Promoted Tweets*. These tweets still have the features of a regular tweet and the users can reply to them, retweet them or bookmark them. The program will have three phases. The first phase will start with a specific small group of advertising companies and it will place tweets that the advertiser has paid for on top of the user's search results. This is actually already happening. The second step will be to provide it also to third-party Twitter clients such as Tweetdeck and Seesmic and split the revenue. The last phase will be to make the promoted tweets available into a user's message stream, even if the user is not following the specific advertiser. The goal of Twitter is to be able to measure whether the tweets resonate with the users and stop them when they do not resonate any more.

4.2 Facebook Business Model

Facebook has introduced several ways of advertising which seem to be quite profitable for Facebook. It has initiated *Facebook pages* and *Facebook social ads*, which are important sources of income. Moreover, Facebook is using virtual money for credits and provides the possibility to shop online through its website.

4.2.1 Facebook Pages

With the help of pages it is possible for companies, products or public personalities to maintain a strong presence on Facebook. The users can become fans of the pages and in this way they become connected with a specific brand or name in general. Every post made by the page will appear on the fans news feed. The important feature is that the owner of the page can measure the engagement and the interaction of the fans through the comments on the page's posts as well as the percentage of news feed read. According to a research [16], fans spend $71,84 more on the brand's products than people that are not fans and they are 41% more likely to suggest the brand to a friend. Moreover, the average value of a fan is $136.38 but of course it can fluctuate.

4.2.2 Facebook Social Ads

Facebook social ads offer a very simple and fast platform for every one interested to promote a product. The platform guide provides the opportunity to select the target group by very precise filtering. The creator of the advertisement can decide for example on general attributes like the age, the gender or the location. The intriguing part is that the advertiser can even specify the college someone attended, the interests and the likes of the users or the groups a

user participates. Those extra criteria can really make the target of the advertising specific and the creator can see in real time how many users fall into his criteria.

For the user, these advertisements appear on the right side banner and the user can interact with them. A strong motive for the user to interact with the advertisement is the list of the friends that have interacted already with it. This list is placed underneath the image.

4.2.3 Facebook credits and Credit Cards

Facebook credits are virtual money that users can use to level up in games and buy items that are not available to the users for free in several applications. The users can buy 10 Facebook credits for $1. 30% of the transactions revenue is earned by Facebook. Facebook also decided to promote credit cards through the social network, which are prepaid cards that a user can buy at a store. The new Facebook cards will be available in values of $15, $25 and $50. More than 150 applications and games are compatible with the use of the credit cards and the goal is to make the cards usable for every Facebook application. The promotion is done with the help of banner advertisements that encourage the users to use these special cards because they offer them reward points for their activities. These prepaid cards were created by GMG Entertainment, which also produced iTunes cards.

4.2.4 Shop Online at Facebook

Facebook has now included a "shop now" tab on brands fan pages through an application called Payvment with an integrated payment system. This idea literally transforms Facebook to s sales platform with 500 million potential customers. All the sales and the progress of the order is done directly from facebook. The shops also offer discounts to the users that "like" their shop or share it with friends.

4.3 Endomondo Business Model

Endomondo is a quite new business. They do not make money through advertising, but they are trying to promote their online shop so that the users will buy their training equipment from the Endomondo site and thus support the social network financially. Moreover, Endomondo has introduce a premium account for business. A business can buy a premium account for every employee, so that they develop healthy and athletic habits. The cost for each employee is 24.85 euros per year, with extra features compared to the free account. In December, 2010 Endomondo also released a pro version priced at $3.99. It offers downloadable routes, headset control, calories counter etc. [18]

4.4 LinkedIn Business Model

LinkedIn provides the users an option for self-advertising called DirectAds. Up to three advertisements can be put on Profile pages, Home pages, the Inbox, Search Results pages and group pages. The advertisements contain a small headline, description, the name of the advertisers and an optional image. Moreover, the advertisement may also contain a URL so that the users can click on it and be redirected on the advertisers website. An important feature of DirectAds is the ability to focus on specific audience. The advertisers can set specific filters according to which the audience of their advertisement will be selected. The criteria can be job function, industry, geography, current company size, seniority, age or gender.

The cost of advertising depends on two parameters. The advertiser can choose to be charged whenever a user clicks on the advertisement (Pay per Click) or pay only for the presence of the

advertisement on several pages (Pay per 1,000 Impressions). The bigger the amount of money the advertiser bids the more likely it is that this advertisement will appear on a page, because there is a competition between the advertisers.

Part of the revenue model of LinkedIn are also the premium accounts for the job seekers as well as for the recruiters. Regarding the job seekers there are three types of accounts beyond the free basic one. These are the Business, the Business Plus and the Executive, priced at $24,95 per month, $49,95 per month, $99,95 per month respectively. They offer more possibilities such as more results, better filters and direct communication with other LinkedIn members. As far as the recruiters are concerned, there are also three types of accounts, each of them giving different possibilities to find a talented employee like full name visibillity and talent filters. The accounts are the Talent Basic for $49.95 per month, the Talent Finder for $99.95 per month and last the Talent Pro for $499.95 per month.

4.5 Foursquare Business Model

Location sharing services are an upcoming trend in social networks these days, since most of the popular internet enabled cell phones contain GPS receivers. Foursquare is getting more and more popular but it is not charging the businesses yet for the services it offers. However, some of the famous brands that made deals with Foursquare did pay an amount of money. The co-founder Dennis Crowley said *"Some deals are paid, some are exploratory, we're all about trying a little of everything and seeing what sticks"*.

4.6 Conclusion

At the beginning of their existence the online social networks had as a goal to broaden their user basis. As soon as the numbers became satisfactory for the creators of the social networks then the goal was to monetize the considerable amount of data available from the users. LinkedIn, Facebook and Twitter are already in the advertising business with each one of them trying a different approach to the advertisers. Furthermore, for example LinkedIn and Endomondo offer subscription accounts with more advanced features than the free accounts. In the future the online social networks are expected to increase their revenue radically as the investors are waiting to take the invested capital back with profit. This might be succeeded by involving more third-party applications. The creators of the applications already pay an amount of money in order to put their application in a social network. It is expected that this amount will rise in the near future, especially the application demands access to personal information.

5 Dangers Of Social Profiling

5.1 Available information

It is important to show what information can be gathered through data mining social networks. The combination of several social networks can lead to a much more complete and thorough profile of a user than one social network alone. The survey is based upon the default privacy settings and the results concern information that is visible when a user is logged in but not a "friend" with the user being observed. Results for the case of not being logged in to a site are not provided for two reasons. First, this case offers very little information and second it is very easy for a malicious user to set up test accounts in every network so that he has better access to every piece of information available. At table 2 we give an insight on the fields that can be collected after combining several social networks and from which network they can be extracted.

The fields of table 2 that are accompanied by a star (*) are mandatory for a user in order to create a profile at the corresponding social network. The rest of the fields are optional and the user is prompted to fill them in while creating the profile. If the user does not fill in the optional fields then several pop ups appear asking the user for more information during the process of profile personalization. The pop ups suggest that the more information the user uploads the easier it is to create connections.

At this point the meaning of some of the table fields should be clarified, so that their functionality and value are better understood. For example *Nickname* for Facebook is the name that the user chooses to appear at the profile page. For Twitter this is an obligatory field and that is the name that appears when a user is browsing the profile of another user. Moreover, the Nickname can be used to address a tweet to a specific user.

Another field with unclear meaning could be the *GPS locations*. Foursquare, Endomondo and Twitter via Twitpic are the networks that expose that kind of information. The photographs that are uploaded by the users in Twitpic contain EXIF [1] data that can be extracted to identify the exact location of the user. The same information can be extracted from Foursquare. In Endomondo the users upload the route of their workout in Google maps so the GPS coordinates of their routine are freely accessible by anyone. *User Content* indicates whether users upload for example photo albums and videos, comment to posts, or post messages themselves. Moreover, the field *Apps bound to other social networks* refers to the binding of accounts of social networks. For example the LinkedIn account can be linked to the Twitter profile of a user so that any update of LinkedIn appears as a tweet in Twitter.

[1] Exchangeable image file format, contains metadata such as date, time, camera settings and geolocation.

	Facebook	Endomondo	Twitter	LinkedIn	Foursquare
First name	Yes*	Yes*	Yes*	Yes*	Yes
Last name	Yes*	Yes*	Yes*	Yes*	Yes
Nick name	Yes(URL)	No	Yes*	No	No
Profile pic	Yes	Yes	Yes	Yes	Yes
E-Mail	No*	No*	No*	Yes*	No
Age	No	Yes	No	Yes	No
Date of birth	No*	Yes	No	Yes	No
Height	No	Yes	No	No	No
Gender	Yes*	Yes	No	No	No
Relationship status	Yes	No	No	Yes	No
City	Yes	Yes	Yes	Yes	Yes
Country	Yes	Yes	Yes	Yes*	Yes
Address	No	No	No	No	No
Mobile Phone	No	No	No	Yes	No
Workplace	Yes	No	No	Yes*	No
Groups	Yes	No	No	Yes	No
Connections	All	Yes	Yes	Yes	Yes
Family Relations	Yes	Yes	Yes	Yes	Yes
Tweets	No	No	Yes	No	No
GPS locations	No	Yes	Yes(Twitpic)	No	Yes
Height	No	Yes	No	No	No
Sport	No	Yes	No	No	No
User content	Yes (wall)	Yes	Yes	Yes	Yes
Apps bound to other social networks	No	Yes	Yes	No	Yes
Search engine	Yes	Yes	Yes	Yes	Yes

Table 2: Data exposed in each online social network.The fields with star (*) are mandatory for the creation of the profile. This information is available only when an account is being used to enumerate the data. Additional information about the values in this table can be found on the previous page.

5.2 User Behaviour

According to a survey [20] it appears that many users are not aware of the danger of exposing their sensitive personal information online. It is interesting to note that 41% of children aged between 8 and 17 had left the default privacy settings at the social network they are using and 44% of the adults admitted that their profile was open to everyone.

Moreover, the research showed that users provide easily personal information and photographs. More specifically, 25% of the users of social networks had posted sensitive personal data about themselves on their profiles. These data included their mobile phone, their address and their email address. The percentage arises when it come to young users, as 34% of them are willing to upload this kind of data. There is also not much consideration when it comes to photographs that could affect someone's reputation or career.

Furthermore, it should be mentioned that users are also willing to accept as friend someone they do not know beforehand. Afterwards, 17% of the adults did talk to the people they didn't know and 35% talked to indirect contacts, such as friends of friends. The fact that some people will add strangers as connections leaves plenty of opportunity to gain access to private data stored on social networks, but this attack vector is beyond the scope of this paper.

5.3 Recent Privacy Incidents

The dangerous effects of exposing sensitive personal information on social networks and especially photographs are already existent, even if the profiles are set to private. Setting a profile to private means that the user can decide that his information is accessible by a certain group of people and not everyone.

5.3.1 Facebook incidents

A recent case from a US court in Pennsylvania shows that not even private profiles are that protected since the court decided to ask for the credentials of the user. The trial was between McMillen and Hummingbird Speedway, Inc. McMillen used to be a driver for the company and had a car crash during a car race. The firm scanned his posts on social networks and came to the conclusion that his injuries were not as serious as McMillen claimed. Then Hummingbird Speedway, Inc demanded from the court the login information of McMillen. They wanted to see the private messages he exchanged with his friends and family on Facebook and MySpace and use them as evidence to the court case if they were related. At first the driver rejected saying that he was protected by the "social network priviledge". The response of the court was that the users should already know that their uploaded data are not private any more since the website operators or any other associated third-party can have access to them. That made the messages non-confidential. Therefore the court ordered McMillen to reveal his credentials to his lawyer so that they could be used for the trial [21]. That shows how risky it can be to upload sensitive information no matter what the privacy settings of the profile are.

During the recent protests in Tunisia, January 2011, people saw Facebook as a tool to upload videos of the real situation in the country as most of the other sites that allowed video uploading had already been blocked by the government. Everyday people as well as activists uploaded videos of the protests and the injured people and passed variable information between them. At the same time Facebook also received many messages about accounts that had been erased or modified without the consent of the owner. Around the Christmas holidays the security team of Facebook realised how serious the security issues in Tunisia were. After a ten-day extensive research the team found out that the Internet Service provider led by goverment commands was running malicious code behind the login page of Facebook in order to acquire all the login

information of the users. The whole country's Facebook credentials were being stolen. This could lead to the direct identification of the activists uploading information on Facebook.

In order to protect the data of the Tunisian users Facebook directed their traffic to an HTTPS server. Furthermore, when users logged out and then directly tried to log in again they were faced with a process of identifying their friends' photos so that they could prove their identity. [25]

Another example is the blackmail of Arab women for money or sexual relationships. They were threatened that revealing pictures of them taken from online social networks would be exposed [22]. According to the police officers most of the times the photographs are the easiest piece of information to extract from a social network and therefore they are most commonly used for blackmailing.

In November 2009 a woman in Quebec, Canada lost her health insurance because of photos she posted. She had declared to her insurance company that she was suffering from severe depression but the pictures she uploaded showed her enjoying her holidays on the beach. That evidence was enough for the insurance company to cut off her allowance [24].

5.4 Twitter incidents

Financial institutions also take advantage of the information the users post online. For example in order to decide whether they should lend money to someone first they scan his online social profile and search for posts that refer to a loss of job or job search. In both cases they assume that the user may have difficulty to pay back the money and pay bills on time so he is a credit risk for the lender [23].

The users should also always keep in mind that their employers may scan their social network profiles. There have been several incidents of people that lost their job positions because of their posts in Twitter [26]. For example, Mike Bacsik, a host radio was fired after making racist comments on Twitter against Mexican people regarding a game. Moreover, Connor Riley had an interview at Cisco and then afterwards wrote on Twitter that she would hate the job but she would definitely appreciate the high salary. Eventually, she was not hired because of this.

There are examples also inside Netherlands. A 39 year old Dutch citizen had to pay a fine of 350 euros for insulting the mayor of Rotterdam through a tweet. The police was monitoring the messages on Twitter and informed the mayor about the offensive content. What exactly was in the message was not made public. [27]

5.5 Combinations of Social Networks

Someone may wonder how can a user's profile from one social network be identified in another social network and therefore gather more information about the user. At this point it should be mentioned that social networks have many properties in common that allow profiles to be matched.

5.5.1 Facebook - LinkedIn - Twitter

In figure 1 is shown which of the information that the users upload in their profile are common between Facebook, LinkedIn and Twitter. All three networks share the first name, the last name and the profile picture.

This means that searching for the full name of a user from Facebook can lead to the right profile in LinkedIn or Twitter. If more than one users appear with this specific full name then there are more properties that can identify the correct user. Between Facebook and LinkedIn the email address, the date of birth and the network of the connections are the keys to exclude

the wrong users and find the match. The location and the exact address can be of much help as well.

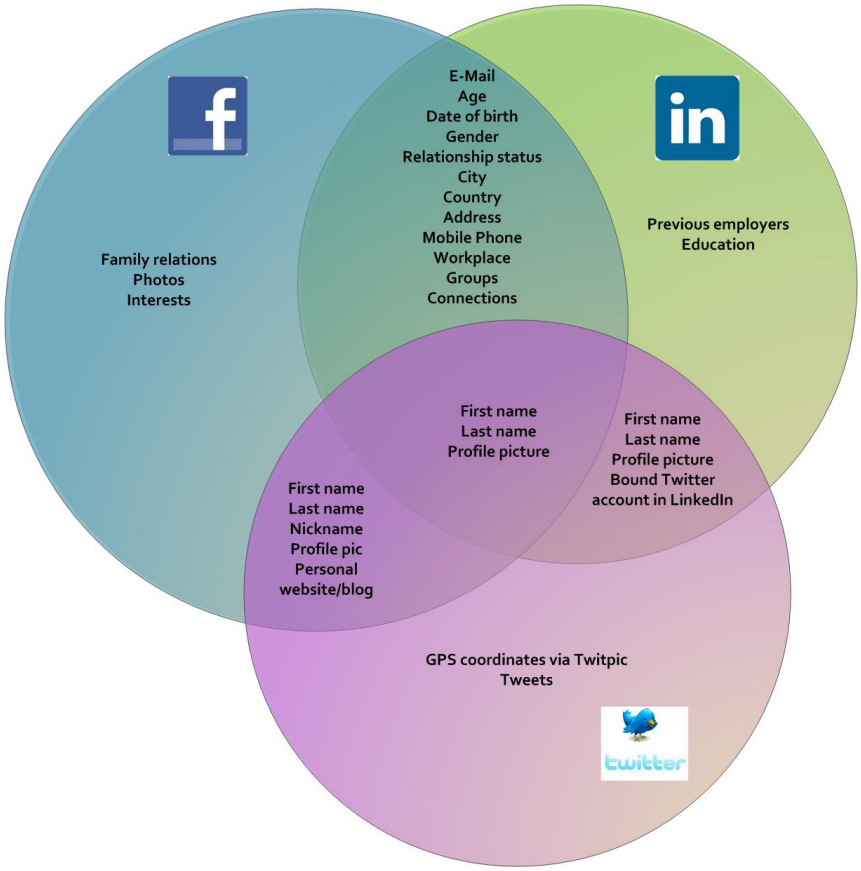

Figure 1: Common properties between Facebook - LinkedIn - Twitter

5.5.2 Facebook - Twitter

As far as the relation between Facebook and Twitter is concerned there we can take advantage of the fact that many users use the same nickname both in Facebook and Twitter and they also post many times their personal website or blog in both social networks. The last clue can lead to 100% certainty about the identity of the user, as this is initiated by the user himself.

5.5.3 LinkedIn - Twitter

There is another strong link between LinkedIn and Twitter. LinkedIn offers the possibility to bind the LinkedIn account to Twitter, which also provides 100% match when someone has access to one of the two profiles at the beginning. Any of the two social networks Facebook and LinkedIn would be a strong point of entry for someone who wants to data mine social networks as they both offer a lot of information standalone but they can be also easily linked to the rest.

5.5.4 Facebook - Endomondo - Twitter

Another worth to mention combination is Facebook, Endomondo and Twitter. Figure 2 provides a graphical explanation of their connections. All three networks share the full name, the profile picture and the date of birth. Besides that, there are some strong links between Facebook

and Endomondo. The strongest one is the direct binding between Facebook and Endomondo. The users of Endomondo can login with their Facebook credentials and every update from Endomondo is directly published at their Facebook profile. If the two accounts are not bound then there is still possibility for identification through the connections network, the gender and the location and then acquire the extra information Endomondo has to offer about the user.

If the available profile comes from Twitter then there is the same advantage of the 100% match because Twitter and Endomondo can also be bound, so that all information from Endomondo appear on Twitter.

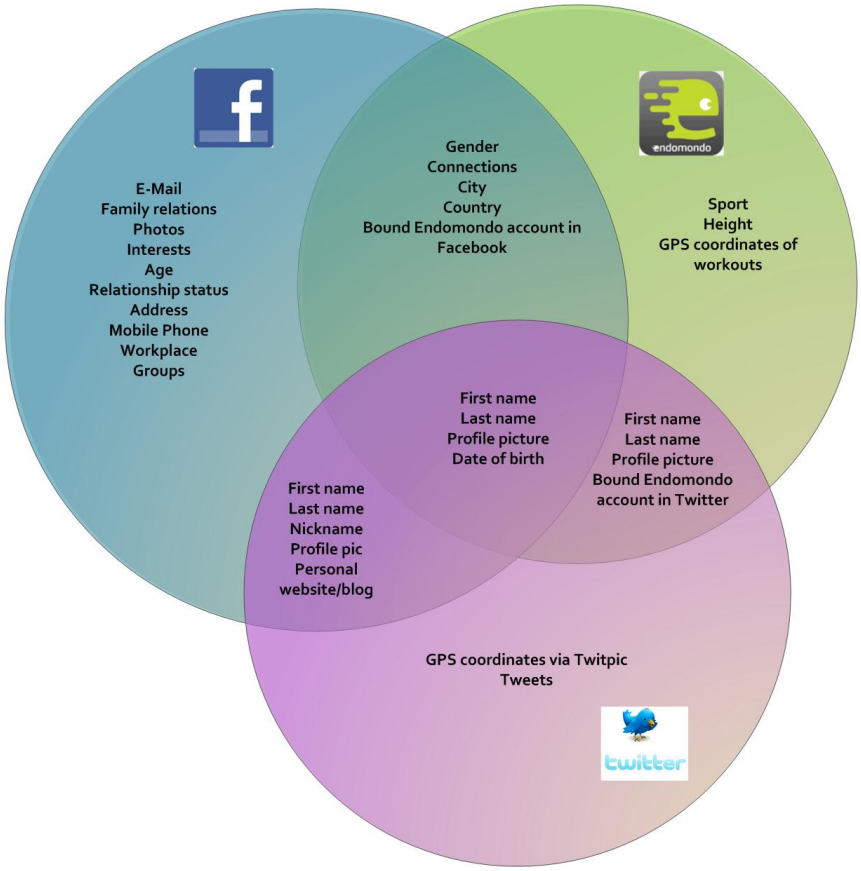

Figure 2: Common properties between Facebook - Endomondo - Twitter

5.5.5 Facebook - Twitter - Foursquare

Another social network that can provide more information about a person's daily routine is Foursquare as it can show the places someone visits. It is significant to find a way to match users from very popular networks such as Facebook and Twitter to Foursquare. The method to match the users is similar to that of Endomondo. Foursqaure in order to become more popular and usable also offers the ability to bind the profile to the one of Facebook and Twitter, as shown also in figure 3.

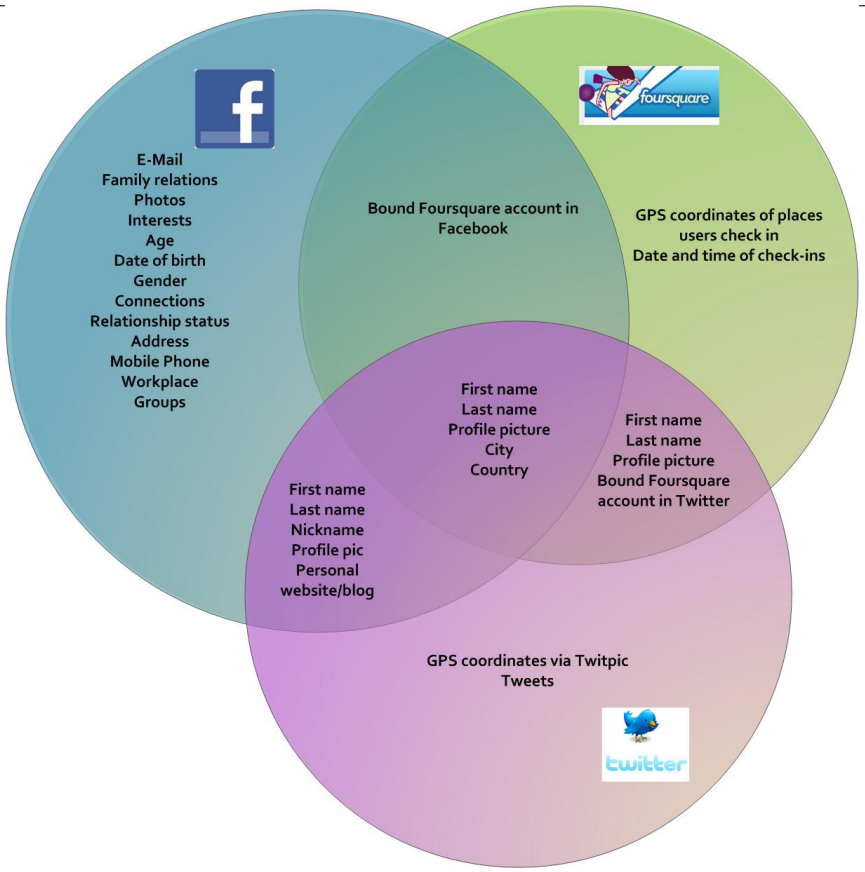

Figure 3: Common properties between Facebook - Foursquare - Twitter

5.6 Risk Scenarios

The examples of the recent privacy incidents show how dangerous it can be to upload thoughtlessly sensitive personal information to an online social network. It is more important to demonstrate that the aggregation of social networks can lead to even more serious privacy violations and risky situations.

5.6.1 Stalking and theft

An example is the combination of information from Facebook and Foursquare or Endomondo. Users of Facebook have the opportunity to state their home address online. At the same time if a user checks in at a venue through Foursquare then any potential attacker is aware that the user is not at home and there is high possibility that the house is empty. That makes the house an easy target to rob without the thief being worried about the owner. There are also users boasting about their new valuable purchases online and therefore making their house an appealing target.

The same effect can have the use of Endomondo. The image of a user's workout is on his profile page and public to anyone. A potential attacker knows in real-time where the person is because the whole route of the workout can be seen in google maps. This means that the attacker can rob the sportsman at a deserted place or rob the empty house.

Another indirect method to find a place someone has visited is Twitpic because it may contain EXIF information. So even if the location of the photo is not recognisable, using the

GPS coordinates one can determine instantly the exact spot the picture was taken.

Moreover, when users check in at a venue using Foursquare then the owner of the venue can identify exactly who the customers are. This first means that he could start targeted advertising by really personalised emails, but there can be more complications. The owner of the venue could also identify the people that accompany the user who checks in. This could cause trouble especially if it was for example a confidential business meeting or a personal meeting and the user did not want to reveal who the other people were. This could cause professional or family implications.

At the chapter about recent privacy incidents it was mentioned that health insurance companies also monitor online social networks. The more networks a person uses the more exposed someone is to the insurance companies. Apart from posts and photographs in Facebook or tweets in Twitter, insurane companies can take a look at the places a user checks in using Foursquare. If a user keeps checking in every single day at a store selling tobacco then this may affect his insurance [38].

Stalkers might use Foursquare and Endomondo to build up a profile of a user's daily routine and know where he is at any time of the day. This could lead to i.e. taking photographs of the target or threatening people.

5.6.2 Identity theft

Fake accounts are also a big issue of the online social networks. There is no control over the data a user will upload in order to create an account. By aggregating the data available in a variety of social networks a malicious person can create a believable fake account and impersonate anyone that uploads enough sensitive personal information. Especially when the fake profile contains the full name, the correct date of birth, the correct workplace and place of living and even photographs of the person it is really difficult to distinguish whether it is a real or fake account. All this data can be collected by scanning Facebook and LinkedIn.

The owner of the fake account can add the friends of the real user. It has been examined that the users tend to accept friend requests of a person they know, even though they are already connected at the social network. Afterwards, the malicious user can exploit the fake account in many ways. For example he can start asking for money pretending he is in a difficult financial position and then give his own bank account number. Friends thinking that their real friend is in trouble might deposit money to the given bank account.

Moreover, through instant messaging friends can make really private conversations on online social networks. A malicious user could take advantage of that and learn really intimate details for a user's life by impersonating their friend using the fake account. This could even include conversations about the professional life of someone and comments on the user's colleagues. Then any kind of blackmail is possible if the user wants those details to remain unknown and prevent the violence of his personal and professional life.

The identity theft can occur of course also outside the virtual world of the social networks. A person that has acquired a handful of information about a user can take advantage of it also in real life, by using a fake name and address or even a fake ID with a photograph downloaded from a social network.

In order to complete the collection of information about a person the malicious user needs also the social security number and the bank account number. These could be acquired by sending a really personalised email impersonating for example the department of Human Resources of the company the user works for and demanding this data because they experience a data loss. The attacker could also try to send an email impersonating large bank institutions for the same purpose.

5.6.3 Law enforcement

It is not only malicious users that monitor online social networks. Police is also very interested in the users profiles and searches for evidence that might indicate criminal behaviour. Users often boast about their crimes online or they are trying to recruit other people. Many times also they use the service of instant messaging or messages to approach especially teenagers. The police can ask for the private information from the social networks about the people that are under investigation.

Every social network has its own policy on the subject. For example Twitter may inform its users before their data is provided to the police unless there is a specific court order not to ask for their conscent. Specifically in the case of Rop Gonggrijp, who was involved in the Wikileaks organisation, Twitter notified him before his data was released to the American Ministry of Justice[37]. For Facebook it is unclear if they notify their users in similar cases. They keep the data of the users for 90 days unless there is a different request from the police [28].

A detailed overview of the policies regarding social networks and law enforcement can be found on the EFF website [42].

5.6.4 Marketing and demographics

Social networking services are indispensable tools for marketing companies. It is possible to build up detailed customer demographics using the information that can be retrieved from social networks, a political party could i.e. gather information on their potential voters. Likewise, a lot of information about the current fans and sympathisants of a particular brand or product. This type of information can already be extracted from the simple proof of concept attached to this report.

5.6.5 Financial criminals

Social networks can provide a wealth of information about an individual. If the information extracted from the social networks can be combined with a social security number and a bank account, then all required information is present in order to acquire a credit card on the targets behalf. In The Netherlands, both missing pieces of information could be looked up by car rental companies, hospitals, the police and employers, which means that there is a very large audience that could get access to this type of information. The information required for an online credit card signup (full name, address, work, ...) can be retrieved from social networks.

Acquiring a creditcard could be done by requesting it from a bank that does not have regional offices and by getting the card delivered to an address of choice. An example of such a bank in The Netherlands is American Express, there are no physical checks of i.e. the ID card involved when a credit card gets requested.

5.6.6 Blackmail

There have been multiple examples of blackmail through social networks in the recent few months [43] [44]. In most cases, these illegal actions get triggered by acquiring data that the victim would rather not disclose publically. The attacker now could try to threaten the victim to meet his/her demands, else the family and friends of the victim will get contacted. This attack could be applied to multiple social networks for numerous scenarios.

5.7 Conclusion

According to table 2 a considerable amount of information about every user can become available so a very detailed profile can be built. The data can be aggregated by combining online social networks. The common attributes that are present in the social networks help to match the users from one network to the other. Sometimes, with for example with Endomondo and Twitter, there can be 100% match if the user links the two accounts.

The available data can be exploited for different purposes. As it is described in the section of risk scenarios the danger of robbery and attack increases rapidly because users share their home address and their real-time location. Potential robbers now can know if the victim is at home and how long time they will have before the owner is back.

Apart from the danger of robberies and attacks, the risks might be more indirect and difficult to detect. Social networks can provide enough information to perform identity theft either online or in the frame of social engineering. This could lead to financial fraud, defamation of the real user as well as personal and professional problems.

Health insurance companies and law enforcement are also monitoring social networks. Health insurance companies try to make sure that their customers are totally honest about their health state and their habits. If they discover different information they can change the funding to the user. On the other hand law enforcement sees social networks as a tool to watch criminal actions. It depends on the social network if it will collaborate with the police and reveal private information about the user.

This chapter also shows examples of the potential new threats that the research team could come up with.

6 Gathering Data

In the previous chapters we described which data can be enumerated and what the relations between the available datafields are. Based on this information, several scenarios can be created to gather and organise this data.

6.1 Enumeration components

The research team was able to construct software to support the following tasks;

- *Facebook - Treeview with connections of targetted user*
  This crawler can create a pedigree for a targetted user. A lot of information about the type of friendship can be derived by building up this tree since it is possible to filter profiles based on (matching) keywords.

- *LinkedIn - Treeview with connections of targetted user*
  The same concept can be applied to LinkedIn. In this way, large parts of company personel can be exposed and mappings can be made for professional relations.

- *Twitter - Twitpic GPS coordinates*
  GPS coordinates can be extracted from Twitpic pictures containing EXIF information. A script was created in order to automate this extraction process.

- *Foursquare - Who was at a venue at certain time?*
  Pictures taken at Foursquare venues display the full names of individual including a link to their profile, a script was made in order to enumerate the information from these photo albums.

Additionally, the following two valuable resources can be retrieved through public access;

- *Twitter; Marketing relevant data*
  By applying keyword analysis, a lot of personal interest can be derived for a single individual. Twitter can provide very specific and up-to-date meta information after the static personal information from an individual has been collected.

- *Endomondo; Plot of user workouts*
  Endomondo offers very detailed workout plots for its users. The GPS data that can be found on a users page can be used to complement a personal profile.

The sourcecode for these tools can be found in the appendix. An updated version of the software will be available at `www.socialsynergy.nl` since the brief research time only allowed to create basic concepts of the tools.

This chapter will explain the workings of the enumerations mentioned above, but first the connections between the social networks are explained in the section below.

6.2 Data mining strategies

There can be multiple incentives to crawl personal data;

- *Crawl targetting one individual*
  This strategy could be used by employers, insurance companies or law enforcement in order to gather as many leads upon a person as possible.

- *Crawl targetting a brand, organisation or company*
  It can be interesting for a company or its competitor to find out who the loyal supporters of a brand or product are. Social networks like Facebook or LinkedIn offer a wealth of related information within an individual personal profile, but the real power can lie in combining the information to build up a partial view of (potential) customers.

- *Crawl everything and process the data afterwards*
  One example of a "complete" crawler is Google, the information in their datastore is built up based on any information they can retrieve [29].

The methods described in this chapter do not exclusively benefit one of these strategies.

6.3 Overview of shared attributes between social networks

After combining the data from table 2, the following mapping of connections between the various social networks was made;

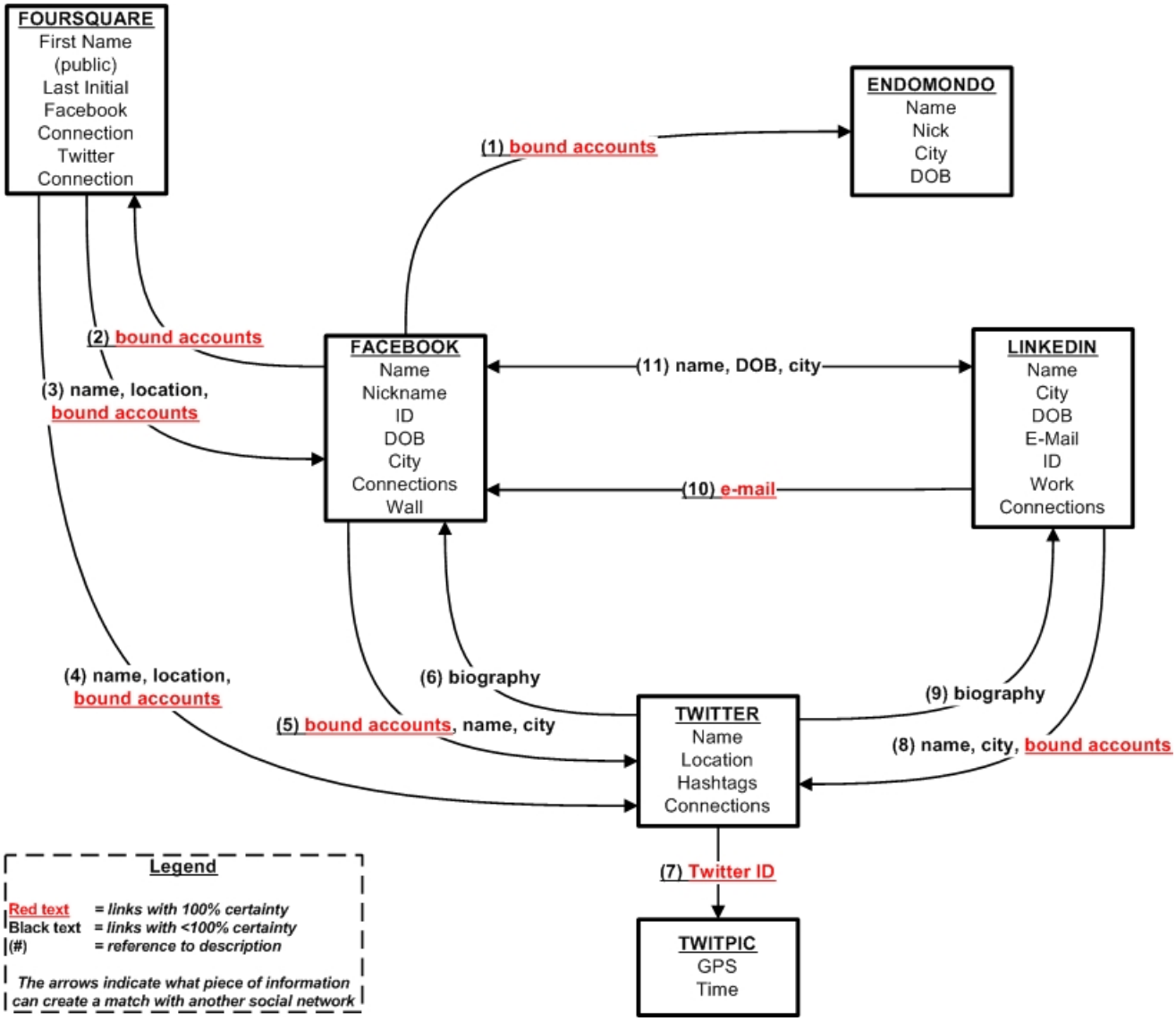

Figure 4: Total overview of datafields shared between social networks

The illustration shows how matches can be established between social networks and what information is required for this. The red and underlined datafields can create a 100% certain match between two social networks since they show a one to one connection between two profiles that a user has. The rest of the matches is harder to make since it is based on values that are not totally unique for a person, like birthday, name or city of residence.

The cells for every social network contain a piece of information that can be retrieved if a match is succesfull. Keep in mind though that these datafields will not always be present, but they should be visible based on the default exposure of profiles. It is also not guaranteed that the data in the fields are correct or true, since users may choose to hide their exposure by obfuscating some of the data.

6.4 Criteria in order to match profiles

The table below gives a description of how a link could be established between two social networks. The line numbers refer to the datafield graphic.

#	How to match the data items?
1	The Facebook "Wall" may contain an application that links directly to someone's Endomondo profile.
2	The Facebook "Wall" may contain an application that links directly to someone's Foursquare profile.
3	The name, nickname and location retrieved from Foursquare may lead to a positive match with Facebook. The account linking option of Foursquare offers a direct link to someone's Facebook account.
4	The name, nickname and location retrieved from Foursquare may lead to a positive match with Twitter. The account linking option of Foursquare offers a direct link to someone's Twitter account.
5	Postings on the "wall" of someone may be put there through Twitter. In this case, the Twitter account name of someone can be retrieved. Additionally, a match can be made based on full name and city.
6	A match between Twitter and Facebook can be made if the biography and name of a Twitter user match to an individual on Facebook.
7	The Twitpic ID of someone is identical to that of Twitter, so this option can always be checked reliably.
8	LinkedIn provides the name and geographical area of a person by default, but a Twitter account can be bound to someones LinkedIn profile as well.
9	A match between Twitter and LinkedIn can be made if the biography and name of a Twitter user match to an individual on LinkedIn.
10	The e-mail of a user is exposed by default on LinkedIn, it can be used on Facebook to retrieve a profile if it was registred with the same e-mail address.
11	The bi-directional connection between Facebook and LinkedIn can be created if the full name, date of birth and/or city match.

Table 3: What data items are needed to create a match and how is this done

6.5 Extending the data based on connections

The section above describes how matches can be created based on user properties, but the connections a user has can be considered one of those properties too. While in some cases a user might have an identical group of friends, acquaintances or connections on multiple networks, it is more likely that a clever matching algorhitmn is needed to determine the overlap. Based on this output, a threshold could be set in order to make the match between two social network accounts definitive. Further exploring the opportunities of this was unfortunatly not possible for the research group because of time constraints.

6.6 Attack vectors when targetting a single profile

Based on the complete diagram of social network components, the following starting points will lead to the best results when building a complete profile about an individual;

1. *Start enumeration on Facebook and go from there* - Facebook has a very open character by default and it maintains a lot of varying information about a person. It will offer the most leads towards matchmaking with other social networks.

2. *Get access to a well-connected LinkedIn account* - LinkedIn accounts can offer a lot of information about company employees, professional experience and they can provide an e-mail for the user.

3. *Find the target on Foursquare* - Foursquare lists the Twitter and Facebook account of a user, which can lead a wealth of information and a confirmed match between the three networks.

6.7 Attack vectors when targetting an interest

Gathering information about this can be best started off by finding a common value between a group of users. The most obvious example is that users are a fan of a certain brand or product on Facebook, but it can also be derived from Tweets or LinkedIn group memberships. The strategy for enumerating this information will vary widely depending on the target though, but it can definitely be done once a common value or property for a usergroup is found.

6.8 Crawlers (Facebook, LinkedIn)

Two data crawlers were built to prove how the datamodel can be filled in as completely as possible. Both of the crawlers start of at a single personal profile and follow the leads that are given there.

1. *Facebook - Treeview of targetted user* - Enumerate the data available to unrelated and not connected friends, this data is public by default.
2. *LinkedIn - Treeview of targetted user* - Enumerate the data available to related connections and associates, since LinkedIn shields connections that are further away than second degree connections.

For the experiments executed, the depth of crawling Facebook was limited three levels deep, as displayed in the diagram below;

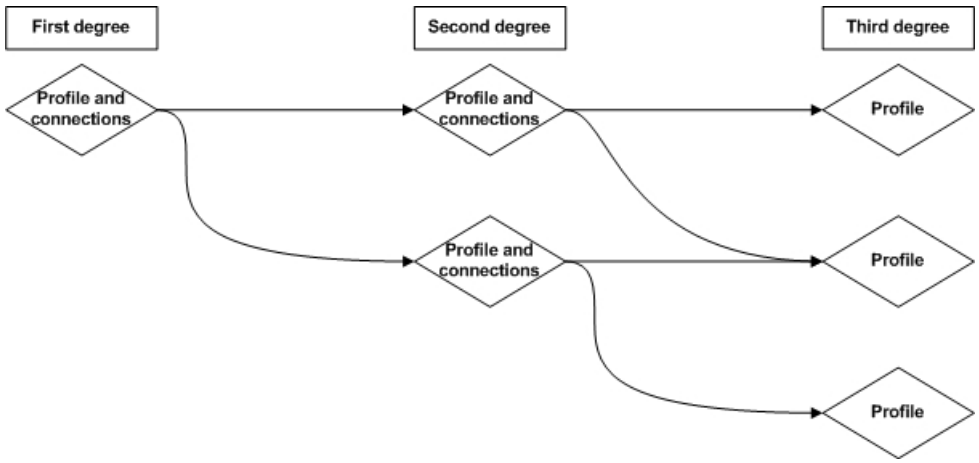

Figure 5: The reach of the crawler programs

Both crawlers can utilize a social network account to gather information normally only available to registered members. In the case of Facebook this means that all information that is not limited by privacy settings could be enumerated over time, just by starting from one well-connected profile. In the case of LinkedIn, an account that gets hacked or where the password gets intercepted could provide a wealth of information for i.e. a competitor. Again, a well connected profile could provide the personal information of several thousands of people that fall within the targets personal network (limited to two degrees). The distinction in information available to the public and members can be found in table 2.

Registered user accounts can be used by the program by sending and receiving cookie strings. These can either be supplied by exporting them from a conventional webbrowser, or by generating them using a "cookie jar". The "cookie jar" is a method to capture and reply with the cookie responses generated by the client and server, so that a webbrowser's behaviour can be emulated.

On a very abstract level, the two programs execute a couple of predetermined steps to reach their output. After supplying the user ID of the person the user wants to enumerate, the program will check if a cookie is present. If not, the user will be asked to enter a username and password after which the cookie is stored. Now that the program can view the same content as a regular user can with a webbrowser, the target's profile can be crawled. From the profile page, a couple of common tags are extracted and stored into the XML file under the users node section.

After this, the target's friend page is enumerated. Since most social networks will not display the connections in one page, these separate pages showing a partial view need to be crawled sequentially. The friends IDs and URLs get stored into a textfile which will be needed by the next step; enumerating the target's friends profile and their connections.

The full abstract set of steps taken by both crawlers is displayed in the illustration below;

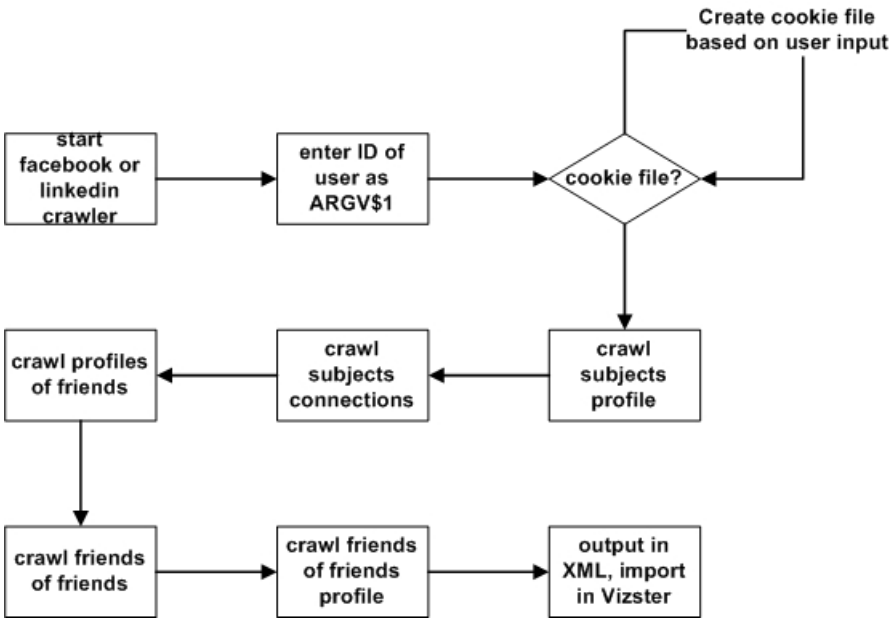

Figure 6: Steps taken by the crawler software

The two social network crawlers utilize the following third party software components;

- PyCurl Python extension - used to supply the cookie and to retrieve the pages crawled using CURL [33]
- Scrapy Python extension - used to extract predetermined tags from the CURL'ed webpages [34]

6.9 Data aggregators (LinkedIn, Foursquare)

6.9.1 Twitpic EXIF extractor

Additionally, a crawler was built to gather picture data from a targetted Twitpic profile. A lot of digital cameras and cellphones add "EXIF" data to pictures and depending on the camera and settings, this may contain information about the date, time and lens settings used for the picture. In addition, some pictures may contain the GPS coordinates of where the picture was taken, which can reveal a lot of extra information about the owner. This data is not displayed in any way on the Twitpic webpage, so it has to be extracted from the pictures themselves.

Further research proved that from the five social networks we researched, only Foursquare and Twitpic leave the EXIF data intact. Facebook removes EXIF tags by default, Endomondo and LinkedIn do not have the option to upload photo albums. An abstract overview of the Twitpic EXIF extractor is displayed below;

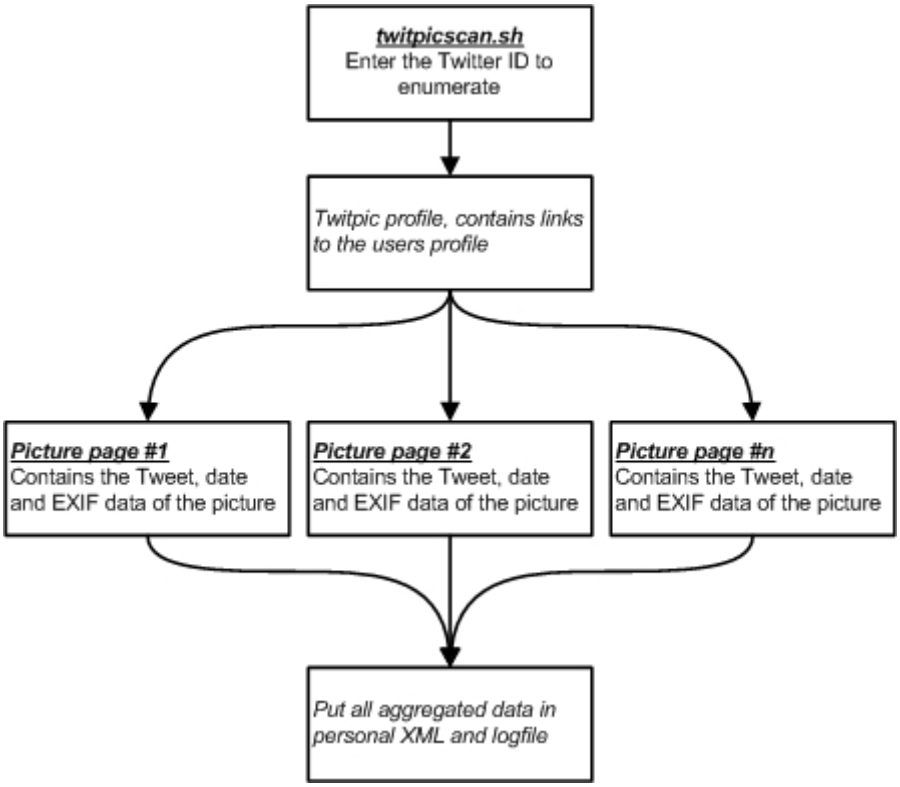

Figure 7: Twitpic enumeration flowchart

6.9.2 Foursquare venue enumerator

Another crawler was built in order to collect information about people taking a picture of a Foursquare venue, so that this information can be used to track who was there and at what time. A Foursquare profile by default does not reveal a lot of data, as seen in table 2. As an example, the profile owners last name is shortened to just the first letter. The full name of the user is listed on picture overviews of a venue though, which means that a list of who took a picture at that point of interest can be created. In addition, the EXIF data of the pictures listed on a venue page can be extracted to derive the time, exact GPS location and sometimes even the full name of the user. An abstract overview of the Foursquare venue enumerator is displayed below;

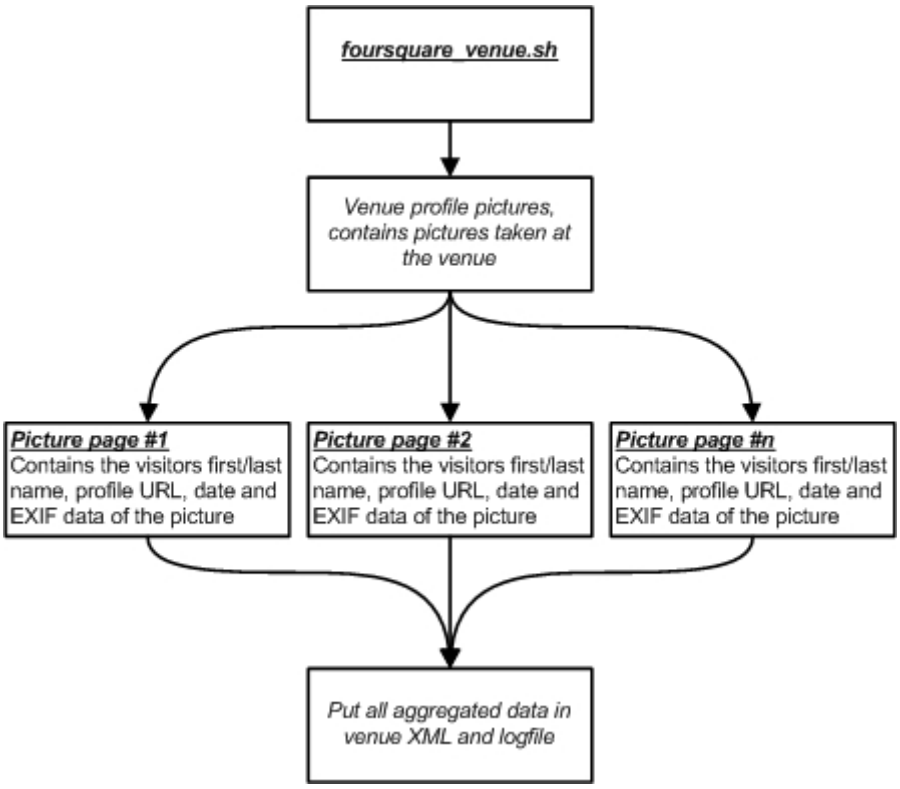

Figure 8: Foursquare enumeration flowchart

6.10 Visualisation

In order to visualize the information we crawl, we have used a slightly modified version of Vizster [35], created by J. Heer of Stanford University [36]. By default, the Java-based program can display the gathered information for a user profile of one social network. It can display highlights between connections and mutual connections and it is possible to visualize data based on various keywords or shared properties of users.

Vizster can use either MySQL databases or XML to display data, by default it comes with a pre-defined set of fields which it will display. In order to display all the information crawled and aggregated, it was necessary to add a couple of custom XML tags to the program, the changed source code and XML layout for this can be found in the appendix.

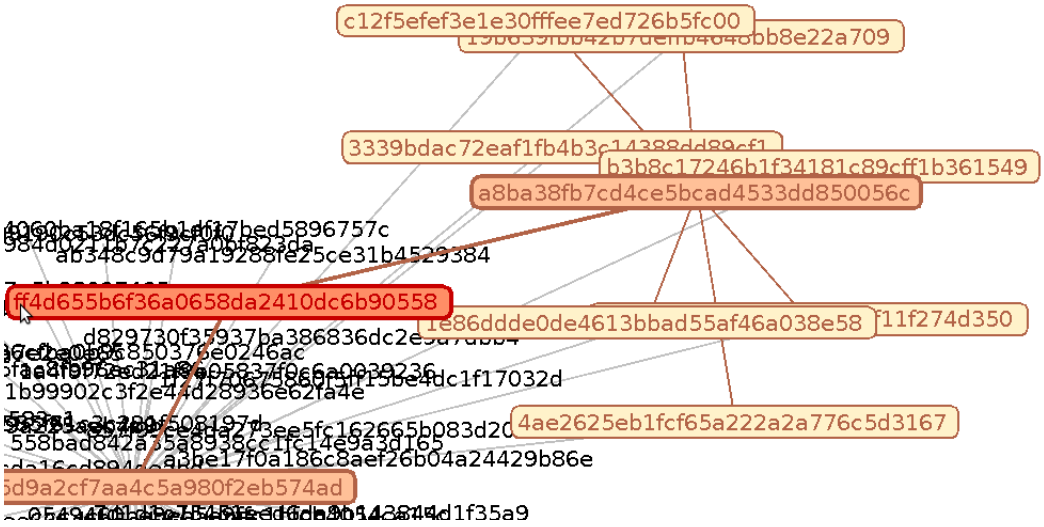

Figure 9: Vizster Example Screenshot (Overview of connections)

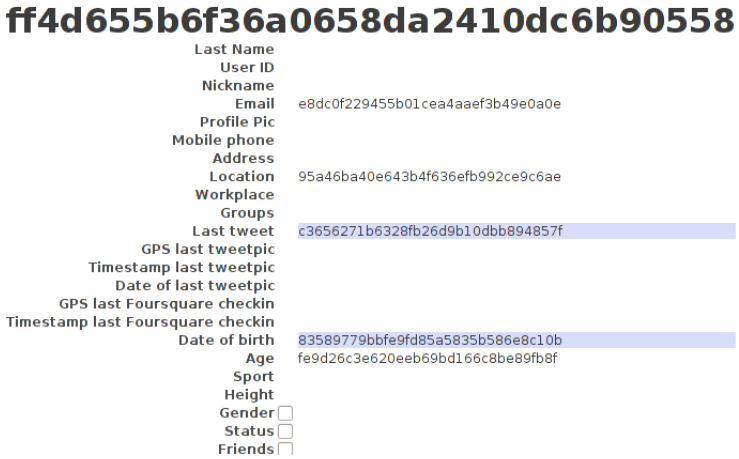

Figure 10: Vizster Example Screenshot (Properties collected)

6.11 XML layout

It was chosen to use XML for the data storage since it is very flexible and easy to debug without the use of additional software. An example of the slightly adapted XML schema can be found below (this schema does not include all possible XML values, but it does offer a good impression of the files layout);

```
1  <graph directed="0">
2   <node id="1">
3     <att name="name" value="Marek"/>
4     <att name="id" value="12345"/>
5     <att name="uid" value="12345"/>
6     <att name="city" value="Amsterdam"/>
7     <att name="dob" value="02/02/2011"/>
8   </node>
9   <edge source="12345" target="54321"> </edge>
10 </graph>
```

7 Countermeasures

The problems stressed in the previous chapters can be mitigated in various ways. There are two categories of countermeasures. First the users themselves could start adjusting the privacy settings and the social networks could take a few measures to prevent data mining. The second category describes the steps that social networks or legislation could take in order to protect the social network users.

7.1 Raising User Awareness

The suggestions in this section could be broadcasted by digital rights lobby groups and companies. These suggestions will benefit the users of social networks since they will develop additional awareness, but they are probably against the interests of the social networks because they could weaken their business models on various points.

7.1.1 Make users aware of what they are "opt-out" to

This suggestion definitely applies to social networks that are expanding their activities by introducing new features. Facebook is the prime example of a social network that by default enables new features for its users, even without an official notification by e-mail or onscreen message [39] [40] [41]. For the other networks, the breaches of privacy appear to be less severe, but this is probably caused by the fact that they are not as data intensive and not that commonly used. The business models described in chapter 4 combined with the exposed data described in chapter 5 give an impression of the risks a user might face.

A countermeassure to this issue could be to include the features a user wants to participate in to the sign-up procedure of the website, or to simply make users aware of the risks. Moreover the users should be informed about the settings they need to change.

7.1.2 Make the users aware of who is interested in their data and why

The datasets that can be created based on social networks could potentially interest many people. While some of the motives for data mining (like i.e. marketing purposes) might appear harmless, others (like i.e. intelligence gathering by governements) might lead to severe consequences for an individual. It is probably best if the users are aware of the implications that participating in a social network might have on their personal life, especially since the developments in this area are happening so rapidly. Chapter 5.5 gives an overview of theoretical scenarios that could harm the user.

7.1.3 Make users aware of data retention on social networks

Data retention can be an issue for individuals as already mentioned in the introductory chapter. While with most social networks it is possible to erase data from the viewable part of a social network, it is hard to say for certain if the data is no longer retrieveable in any form (either personal or anonymous). Another risk can be caused by the fact that third parties can take snapshots or backups of certain datasets as proven in some of the proof of concepts. Once the data has been stored on third party machines, it is virtually impossible to erase it completely. The analysis of social networks in chapter 3 gives a brief overview of how the social networks claim to handle data.

7.1.4 Make users aware of EXIF data in phone cameras

By using the Twitpic enumeration code provided in the proof of concept, it has been shown that the GPS coordinates supplied by EXIF can be enumerated for a specific Twitpic user. The combination of the GPS coordinates, the time, the Twitter username and the picture means that a plot of the user's most recent positions can be created and imported into i.e. Google Earth/Maps. Again, the most effective countermeassure would be to educate users about the consequences this functionality could have, especially if a personal profile can be built up over a longer period of time.

7.2 Restricting Access To User Data

The recommendations given in this section are aimed towards the social networks themselves. The research group has been able to exploit the openness of some social networks without any direct consequences. These possibilities are harmfull for both the users and the social networks.

7.2.1 Limit the exposure of a contacts relations

As seen in the previous chapter, it would theoretically be possible to enumerate the whole of Facebook and large parts of LinkedIn right now. Twitter could be fully enumerated by default while Endomondo and Foursquare have the data shielded of fairly well from connections that are not directly related to the "target". Especially in the case of Facebook, it would be very wise to limit the exposure connections a person has to i.e. people with mutual friends. There is no valid or practical reason to expose this data by default to unrelated users.

The model used by LinkedIn (which displays personal data up to two degrees away) is already an improvement, but it can still offer a very large partial view if the node from which enumeration starts has many connections. Large parts of a companies personnel list can still be enumerated, which might be disturbing for some organisations, especially if this data gets matched with their employees personal data.

Twitter follows a completely different approach by leaving all the users connection out in the open, but this is not necasserily a bad thing since their keyfunction is offering a very open messaging service. It also does not offer a comprehensive overview of a users personal data, although a lot of personal data could be derived from a users tweetstream and the hashtags used there.

Twitter has recently started to (optionally) include a user's GPS coordinates into Tweets, which is information that can be enumerated, analyzed and processed as well. The accounts a Twitter user follows and his own followers are exposed by default. This information could be cross-checked with other social networks in order to build up a personal profile. It has to be mentioned that the connections between users on Twitter are not always mutual, since it is possible to "follow" someone without "following" them back.

7.2.2 Do not disclose truly unique identifiers

Truly unique identifiers can include e-mails or nicknames. If this data can be extracted from a social network (either through a URL or from profile information), then the reliability of the matching algoritmns can be increased substantially.

The same applies to bound accounts, which are the links between social networks. While this feature might be useful for a connection, it is probably unnecassery to display the linked social networks to the public.

7.2.3 Block accounts with excessive data access volumes

The research group has been able to enumerate personal data from social networks using the programs covered in the proof of concept. Multiple testruns have been executed using the same test accounts and IP addresses without any consequences at all, even though the volumes of data retrieved were much higher than an average "legit" user would require. This means that a third party with malicious intentions could start doing the same without facing any direct obstructions.

There might be some rate limiting features present in the social networks that have been crawled. However they might be put into force if more information gets enumerated. With their current settings the research group was already able to enumerate multiple degrees of connections for a user.

8 Conclusion

Our conclusion to the main research question has been derived by answering first the subquestions. The same pattern will be followed in this chapter as well. First the partial conclusions from the subquestions will be quoted and afterwards the conclusion for the main research question.

8.1 Meaning of online privacy

The first subquestion concerns the meaning of online privacy. Although people claim their right to protect their personal information in their everyday life when it comes to online social networks they do not realize to what extend their data is unprotected. The rapid development of social networks has given no time to people to actually think if these networks violate their privacy and whether they should be used. It is important to stress out to the users that protecting their privacy means avoiding pitfalls.

8.2 Exposure of social networks - Privacy statements

The default settings of social network profiles allow a great deal of information to be public. As it is shown in table 2 a full profile can be built about a user without the need of a direct connection with him. The most striking information that can be collected is the home address, the mobile phone, the workplace as well as the GPS coordinates of the uploaded pictures.

It appears that the social networks leave to the users the responsibility to change the default privacy settings. However, this is not part of the signing up procedure.

The privacy policies of the social networks contain some controversial statements. The most striking statement is the anonymization of the data. Facebook, Endomondo and LinkedIn anonymize the user data before distributing them. However, the algorithm for the anonymization is not mentioned so it is not certain if the data is effectively anonymized. Moreover, it is not stated who they distribute the data to. Moreover, this data is used for personalised advertising and emails.

Furthermore, each social network gathers as much data as possible for its users. For example Facebook collects data about user-behaviour from third-party websites using the “like” button and cookies.

8.3 Business models associated with privacy

The business models that the social networks follow are also part of the conducted research, especially because they can be related to the privacy of the users. The social networks are trying to monetize the information they have acquired through the users’ uploads. Facebook, LinkedIn and Twitter are already in the advertising business with personalised advertising and banner advertisements.

The social networks offer filtering services to the advertisers so that they can target specific population. The filtering can be based on gender, age, location, education etc. There are no official figures about the exact profit of the social networks. Subsequently, there is no evidence that they directly sell the users’ data but it is possible that they are already doing it. Premium accounts can also be a source of income. For example LinkedIn and Endomondo have paid accounts with more advanced features.

8.4 Combination into something dangerous

The synergy of social networks can be based on the fact that they have many attributes in common. In the figures 1, 2 and 3 it is visible which fields are shared between the networks and therefore can be used for user matching. The attacks against the users of social networks can include robbery and stalking if the home address is exposed in combination with real-time GPS location of the user (Facebook - Endomondo - Twitpic).

Apart from that, another danger would be identity theft (LinkedIn - Facebook). This could either be performed in the frame of the social networks or for social engineering. Identity theft can lead to financial fraud and personal misunderstandings. One should not forget that if his current workplace is exposed then it could lead to professional issues as well.

Health insurance companies attempt to locate users that might hide information about their health state. If the insurance company discovers evidence on a social network that does not match the client's data they can prosecute the client and request money. Moreover, law enforcement is searching for illegal or activist activity on the social networks and there are cases that people have been fined or prosecuted over their posts.

Blackmail is another serious danger that can be derived from careless usage of online social networks. The uploaded photographs are the main "weapon" of the blackmailers. They can download and save photographs on their hard drive and use them to blackmail someone over money for example. Apart from the above, it should be mentioned that companies that issue credit cards try to locate their debtors on online social networks and try to embarass them to their connections by sending messages about their debts.

8.5 Attack vectors

According to our research Facebook, LinkedIn and Foursquare can be the starting points for an attack to gather information from a variety of social networks. It is easy to build crawlers especially for Facebook that is quite open. As it can be seen in the proof of concept when starting from one user at Facebook a tree can be built with the connections information and their profiles. The same method can be used for LinkedIn. Moreover, a crawler for Twitpic was built to prove how easy it is to extract the EXIF data from photographs. The same can be achieved for Foursquare. All this data gathered can be used to match profiles from one network to another based on specific common characteristics. This can be seen at the table 3

8.6 Countermeasures

Raising user awareness is the main way to confront the dangers of social profiling. The users should be informed about they can opt-out from the applications that are enabled by default. Moreover, it is of great significance to inform them about the fact that when their data goes online then it stays online. Even if the data is erased from the online social network it can appear on other websites or be saved on someone's hard drive. There are many third parties that could be interested in the users' data so it is advisable to let the users know that. The last issue that conceerns the users is the EXIF data contained in photographs. It would be advisable for the users to remove this before they upload a photograpgh to a social network.

It is not only the users that should take measures against data mining social networks. The social networks should restrict access to crucial attributes such as connections and emails. It would also be a good idea to block accounts that try to access an excessive amount of user data.

8.7 What are the privacy risks associated with social network user profiling?

As we can conclude from the subquestions the risks have many parameters. They are associated with the business models of the social networks and with third parties that try to exploit the users' data. The risk scenarios that we finally came up with are theft, stalking, identity theft and blackmail. Apart from that, monitoring from from insurance companies, the police, credit card companies or corporate spies can prove to be also dangerous.

9 Appendix

The appendix holds the simple proofs of concept that were built during the runtime of the project. Updated and better performing versions of these programs will be posted on `www.socialsynergy.nl`.

9.1 Twitpic EXIF extractor

The script that extracts GPS coordinate for any Twitpic user and outputs a KML file useable by Google Earth and Google Maps.

```
1  #!/bin/bash
2  #
3  # Marek Kuczynski
4  # marek.kuczynski@os3.nl
5  # www.socialsynergy.nl
6  #
7  #
8  # use twitter user name as argument $1,
9  # (optional) destination directory as $2
10 #
11 # created this quick and dirty, feel
12 # free to improve it!
13
14 echo "enumerating $1 on `date`"
15
16 # remove and recreate previous files
17 rm $2$1.kml &>/dev/null
18 rm $2$1.full &>/dev/null
19 touch $2$1.kml
20 touch $2$1.full
21
22 # make the header for the KML file
23 printf "<?xml version=\"1.0\" encoding=\"UTF-8\"?>\n<kml xmlns=\"http://earth.
       google.com/kml/2.0\">\n<name>Twitpics by #marekq</name>\n<open>1</open>\n<
       Folder>" >> $1.kml
24
25 # check how many twitpic pages there are for the user
26 pagenumb=`curl -s http://twitpic.com/photos/$1?page=9999 | egrep -o '\?page
       \=[0-9]*\"' | sed s/\?page\=//g| tr -d '"'`
27
28 # add one to get the total amount of pages
29 pagenumb=$(( $pagenumb + 1 ))
30
31 # start crawling the pages from pagenumb till 0
32 while [ $pagenumb -gt 0 ]
33 do
34         # get the list with URLs from the twitpic profilepage and store them in a
               temporary file
35         curl -s http://twitpic.com/photos/$1?page=$pagenumb | egrep -o '<a href="
               /.{6}">' | cut -c 11-16 >> $1.p$pagenumb
36
37         # enumerate the url's from the tempfile and leech the images
38         for i in `cat $1.p$pagenumb`
39         do
40
41                 # wget the page containing the photo
42                 wget -q -O w$i http://twitpic.com/$i
43
```

```
44                  # extract the actual picture URL (theyre either large or fullsize
                        pics) and the Tweet attached
45                  wget=`cat w$i | egrep -o 'src\=\".*photos/large.*\" alt'\|'src\=\
                        ".*photos/full.*\" alt' | awk '{ print $1 }' | cut -c 6-999 |
                        tr -d '"'`
46                  tag=`cat w$i | egrep '<title>.*</title>' | sed s.\<\/title\>..g |
                        sed s.\<title\>..g | sed s.on\ Twitpic..g`
47
48                  # download the first 30k of the pic (usually is enough to get the
                        EXIF) and run the exif-tool on it
49                  curl -r0-50000 -s -o pic.jpg $wget
50                  picgpsposition=`exiftool pic.jpg -GPSposition`
51
52                  case "$picgpsposition" in
53
54                  *GPS*)
55                          picday=`exiftool pic.jpg -GPSdatestamp | awk '{ print $5
                                }'`
56                          picdate=`exiftool pic.jpg -GPStimestamp | awk '{ print $6
                                }'`
57                          picgpslat=`exiftool pic.jpg -GPSlatitude`
58                          picgpslong=`exiftool pic.jpg -GPSlongitude`
59                          picgpsatt=`exiftool pic.jpg -GPSaltitude | awk '{ print
                                $4,$5,$6,$7,$8 }'`
60
61                          # convert GPS coordinates from minutes to decimals (basic
                                maths, hell yeah)
62                          latone=`echo $picgpslat | awk '{ print $4 }'`
63                          lattwo=`echo $picgpslat | awk '{ print $6 }' | tr -d "'"`
64                          latthree=`echo $picgpslat | awk '{ print $7 }' | tr -d '"
                                '`
65
66                          longone=`echo $picgpslong | awk '{ print $4 }'`
67                          longtwo=`echo $picgpslong | awk '{ print $6 }' | tr -d "'
                                "`
68                          longthree=`echo $picgpslong | awk '{ print $7 }' | tr -d
                                '"'`
69
70                          # make a east coordinate negative
71                          longcompas=`echo $picgpslong | awk '{ print $8 }' | tr -d
                                'E' | tr "W" "-"`
72
73                          # make a south coordinate negative
74                          latcompas=`echo $picgpslat | awk '{ print $8 }' | tr -d '
                                N' | tr "S" "-"`
75
76                          # do the conversion
77                          calclat=`wcalc -q -P4 $latone+\($lattwo/60+$latthree
                                /3600\)`
78                          calclong=`wcalc -q -P4 $longone+\($longtwo/60+$longthree
                                /3600\)`
79
80                          # get the coordinate string
81                          coordinate=`echo $longcompas$calclong, $latcompas$calclat
                                `
82
83                          # build the KML record
84                          echo found coordinate!
85                          printf "<Placemark>\n<description>$tag $picdate $picday</
                                description>\n<ImageDescription>$tag $picdate $picday
```

```
                                </ImageDescription>\n<name>$tag $picdate $picday</name
                                >\n<Point>n<altitudeMode>clampedToGround</altitudeMode
                                >\n<coordinates>$coordinate</coordinates>\n</Point>\n
                                </Placemark>\n" >> $2$1.kml
86
87                                  # export a kml file (EXIFtool export, lame!)
88                                  # exiftool -p kml.fmt pic.jpg > $2$1.$picday.
                                        $picdate.kml
89
90                          # echo a string with the stuff found to the 'full' file
91                          # printf "<node id="1">\n<att name="$1">"
92                  ;;
93
94                  *)
95                          echo no coordinates found in $i, \( $wget \)
96                  ;;
97
98                  esac
99
100                 rm w$i
101         done
102
103         # clean up
104         rm $1.p$pagenumb
105         pagenumb=$(( $pagenumb - 1 ))
106 done
107
108 # delete the picture
109 rm pic.jpg
110
111 # close the KML file
112 printf "</Folder>\n</kml>" >> $1.kml
113 exit 1
```

9.2 Foursquare Venue Enumerator

Enter a venue ID in order to see who took a picture there.

```
1  #!/bin/bash
2  #
3  # Marek Kuczynski
4  # marek.kuczynski@os3.nl
5  # www.socialsynergy.nl
6  #
7  #
8  # use a foursquare venue ID as argv1
9  #
10 # created this quick and dirty, feel
11 # free to improve it!
12
13 #download venuepage including pictures according to user input ($1). save it as
       venue#.pid
14 curl -s http://foursquare.com/venue/$1/photos | egrep -o 'view_photo\?id=.{24}' >
       web
15
16 #for every picture CURLed...
17 for i in `cat "web"`
18 do
19         #...download the photopage to retrieve the actual full size pic and the
               picturetakers name
```

```
20          curl -s -o w_$i http://foursquare.com/$i
21          #...strip down to get the URL of the fullsize picture and download the
                pic
22          wget=`cat w_$i | egrep -o 'src=".*jpg' | sed 's/src="//g' | sed 's/"//g'`
23          wget -q $wget -O p_$i
24          #...process the picture to retrieve EXIF date and time information and
                store it
25          picday=`exiftool p_$i -FileModifyDate | awk '{ print $5 }'`
26          picdate=`exiftool p_$i -FileModifyDate | awk '{ print $6 }'`
27
28          #...retrieve who uploaded this picture from the webpage, grab the user id
29          usernickname=`cat w_$i | egrep -o '="\/.*\"' | tr -d '\" \/ \=' |  sed s
                /^user//g`
30
31          #debug
32          #echo $usernickname >> nicks.txt
33
34          #...retrieve who uploaded this picture from the webpage, grab the full
                user name (yes, this is privacy sensitive!)
35          userrealname=`cat w_$i | egrep -o "$usernickname\"\>.*\<\/a | sed s/
                $usernickname//g | tr -d '\"\>\<' | sed 's/\/a//g'`
36
37          #append the info we were able to gather to a comma seperated string we
                can store in a db
38          echo $usernickname,$userrealname,$picday,$picdate,
39  done
```

9.3 Facebook Profile Crawler

9.3.1 Facebook Main Program

Used to issue commands to the library file. This program is in the process of being rewritten to fully Python, it is fully functional however. Check `www.socialsynergy.nl` soon!

```
1   #!/bin/bash
2   #
3   # Marek Kuczynski
4   # marek.kuczynski@os3.nl
5   # www.socialsynergy.nl
6   #
7   #
8   # library used for enumerating Facebook
9   # grab the profile of the ID given in $1
10  #
11  # created this quick and dirty, feel
12  # free to improve it!
13
14  # store the targets name for future use
15  target=$1
16
17  # (re)create the XML file using the name of the crawling target
18  printf "<graph directed=\"0\">\n" > r$target.xml
19
20  ### ENUMERATE FUNCTION
21
22  enumerateprofile (){
23          if [ -f $1.pdump ]
24          then
25                  echo $1\'s profile already enumerated, skipping
26          else
```

```
27                  # dump the profile information to a temporary file
28                  python fb_lib.py p $1 > $1.pdump
29
30                  # grab some interesting data from the pagedump, to be extended
31                  username=`cat $1.pdump | grep -o 'profile_name.*</strong>' | sed
                        's/strong\>//g' | sed 's/profile_name//g' | tr -d '[",>,<,/]'`
32                  id=`echo $1`
33
34                  # create a record for the user we're enumerating, push it to
                        r$target.xml (r == root)
35                  printf "<node id=\"$1\">\n" >> r$target.xml
36                  printf "   <att name=\"name\" value=\"$username\"/>\n" >>
                        r$target.xml
37                  printf "   <att name=\"id\" value=\"$id\"/>\n" >> r$target.xml
38                  printf "   <att name=\"uid\" value=\"$1\"/>\n" >> r$target.xml
39                  printf "</node>\n" >> r$target.xml
40          fi
41  }
42
43  enumeratefriends (){
44          # write the ID of the enumerated friendpage to a file so that we don't
                scan it again
45          if [ -f $1.fdump ]
46          then
47                  echo $1\'s friends already enumerated, skipping
48          else
49                  # calculate the amount of friends
50                  friendsnr=`cat $1.pdump | egrep -o 'Friends\ \(.{0,10}\)' | tail
                        -1 | tr -d [F,a-z,\>,\',\",\(,\)]`
51
52                  # calculate the amount of pages that need to be crawled
53                  pages=`echo $friendsnr | sed 's/.$//'` && pages=$(( $pages + 1 ))
54
55                  # now we have enough info to call the enumerate option of the
                        library, it will dump the HTML of all friendpages for the
                        target
56                  python fb_lib.py e $1 $pages > $1.fdump
57
58                  # extract the ID's and nicknames from the friends dump file and
                        put them in $1.idlist
59                  cat $1.fdump | egrep -o 'connect.php\?id\=[0-9]{6,16}'\|'profile.
                        php\?id\=[0-9]{6,16}' | tr -d [a-z,?.=,.] | sort | uniq > $1.
                        idlist
60
61                  # generate the edges XML file with edges for Vizster. It'll be
                        merged with the mail file in the end to keep the edges at the
                        bottom of the file
62                  # the ID's originate from the $1's profile and are written to the
                         $target edges XML file
63                  for i in `cat $1.idlist`
64                  do
65                          printf "<edge source=\"$1\" target=\"$i\"> </edge>\n" >>
                                edges.$target.xml
66                          echo $i >> hitlist.$target
67                  done
68          fi
69  }
70
71  ### EXECUTE SECTION
72
```

```
73 # start of by enumerating the ID given on the commandline ($target => many)
74 enumerateprofile $target
75 enumeratefriends $target
76
77 # after this, run the enumeration on the targets friends ($target friends => many
       )
78 # second degree connections of the target will be mined for friends and their
       profile info
79 for i in `cat $1.idlist`
80 do
81         enumerateprofile $i
82         enumeratefriends $i
83 done
84
85 # now, mine the second degree connections personal data, but not the connections
       (unless you really want to enumerate the complete facebook ofcourse :-) )
86
87 for i in `cat hitlist.$target`
88 do
89         enumerateprofile $i
90 done
91
92 # merge the edge information to the main file
93 cat edges.$target.xml >> r$target.xml && rm edges.$target.xml
94
95 # close the file
96 printf "</graph>" >> r$target.xml
```

9.3.2 Facebook Library File

Used to retrieve Facebook pages by using a cookie. Stores all downloaded data on disk as a cache.

```
1  # Marek Kuczynski
2  # marek.kuczynski@os3.nl
3  #
4  # library used for enumerating facebook
5  # use it combined with fb_exec.sh
6  #
7  # sudo apt-get install python-pycurl \
8  # python-dev libcurl4
9  #
10 # OPTION 1 what def to call
11 # OPTION 2 friend ID to enumerate
12 # OPTION 3 max amount of friends
13
14 import os, sys, re, pycurl, os.path, StringIO
15 string = StringIO.StringIO()
16
17 ### DEFINITIONS
18
19 def __init__(self, username, password, friendid, pagenr):
20         self.username = username        # text format, so i.e. marek.k
21         self.password = password        # regular password to login to fb
22         self.friendid = friendid        # number format, so i.e. 1018929754
23         self.pagenr = pagenr            # fb displays 10 friends/page, choose a #
               to start with after which a 0 is added (so 5 => 50)
24         self.url = url
25
26 # CURL SETUP in order to communicate using a cookie
```

```
27
28 def createCookie(username, password):
29         login = pycurl.Curl()
30         login.setopt(pycurl.URL, "https://login.facebook.com/login.php?m&amp;next
               =http%3A%2F%2Fm.facebook.com%2Fhome.php");
31         login.setopt(pycurl.FOLLOWLOCATION, 1);
32         login.setopt(pycurl.POST, 1);
33         login.setopt(pycurl.SSL_VERIFYPEER, 0);
34         login.setopt(pycurl.POSTFIELDS, "non_com_login=&email=" + username + "&
               pass=" + password);
35         login.setopt(pycurl.ENCODING, "");
36         login.setopt(pycurl.COOKIEJAR, "fb.cookie");
37         login.setopt(pycurl.USERAGENT, "Lynx/2.8.8dev.3 libwww-FM/2.14 SSL-MM
               /1.4.1");
38         login.perform()
39         print "cookie set!"
40
41 # ENUMERATE FUNCTION, crawls a page based on the input URL
42
43 def enumerateFriend(url):
44         retrieve = pycurl.Curl()
45         retrieve.setopt(pycurl.URL, url);
46         retrieve.setopt(pycurl.FOLLOWLOCATION, 1);
47         retrieve.setopt(pycurl.COOKIEFILE, "fb.cookie");
48         retrieve.setopt(pycurl.COOKIEJAR, "fb.cookie");
49         retrieve.perform()
50
51 ### CHECK FOR COOKIE, else create it
52
53 if os.path.isfile("fb.cookie"):
54
55         print "cookie found! in case the script doesn't run properly, recreate it
               by deleting fb.cookie from " + os.getcwd()
56
57 else:
58
59         print "a cookie was not found at " + os.getcwd() + ", enter your username
               (i.e. \"marek.k\") and password to recreate it"
60         user = raw_input("enter username...")
61         pasw = raw_input("enter password...")
62         createCookie(user, pasw)
63
64 ### PROGRAM OPTIONS, call with E or P as argument 1
65
66 if sys.argv[1] == "e": # enumerate all friendpages
67
68         for i in range(0, int(sys.argv[3])):
69
70                 print "\n ------------ PAGE " + str(i) + " ------------\n"
71                 enumerateFriend("http://m.facebook.com/friends.php?id=" + sys.
                       argv[2] + "&f=" + str(i) + "0&refid=5&ref=pymk")
72
73
74 if sys.argv[1] == "p": # retrieve friendpage
75
76         enumerateFriend("http://m.facebook.com/profile.php?id=" + sys.argv[2] + "
               &refid=7")
```

9.4 LinkedIn Profile Crawler

Not stable yet, check `www.socialsynergy.nl` soon!

9.5 Vizster Changed Sources

Modify the following lines at the end of the ProfileClass.java file located in '/src/vizster/ui/';

```
    public static final String[] ATTR = {
        "name",
        "last_name",
        "uid",
        "nickname",
        "email",
        "profile_pic",
        "mobile_phone",
        "address",
        "location",
        "workplace",
        "groups",
        "last_tweet",
        "last_gps_twit_pic",
        "last_gps_time_twitpic",
        "last_gps_date_twitpic",
        "last_gps_4sq",
        "last_gps_time_4sq",
        "dob",
        "age",
        "height",
        "sports",
        "gender",
        "status",
        "nfriends",
    };

    public static final String[] LABEL = {
        "Name",
        "Last Name",
        "User ID",
        "Nickname",
        "Email",
        "Profile Pic",
        "Mobile phone",
        "Address",
        "Location",
        "Workplace",
        "Groups",
        "Last tweet",
        "GPS coordinates from last tweetpic",
        "Timestamp of last tweetpic",
        "Date of last tweetpic",
        "GPS coordinates from last Foursquare checkin",
        "Timestamp of last Foursquare checkin",
        "Date of birth",
        "Age",
        "Sport",
        "Height",
        "Gender",
        "Status",
        "Friends",
    };
```

9.6 Bonus - Create A List Of The 1000 Most Followed Twitter Accounts

This script builds a list of the 1000 most followed Twitter accounts worldwide, which can be a good starting point for crawling or EXIF GPS data.

```
1  #!/bin/bash
2  #
3  # Marek Kuczynski
4  # marek.kuczynski@os3.nl
5  # www.socialsynergy.nl
6  #
7  #
8  # run the script to get a textfile with the most
9  # followed tweeters
10 #
11 # created this quick and dirty, feel
12 # free to improve it!
13
14 for i in `seq 100 100 1000`
15 do
16         curl -s http://twitaholic.com/top$i/followers/ | egrep 'statcol_name' |
               egrep -o '"/.*/"' | tr -d '"/' >> celebs.$$
17 done
18
19 sort celebs.$$ > celebs_sort.twit
20 rm celebs.$$
```

List of Tables

List of Figures

References

[1] *Facebook Wikipedia, January 2011*
`http://en.wikipedia.org/wiki/Facebook#Criticism`

[2] M. Balduzzi , C. Platzer, T. Holz , E. Kirda, D. Balzarotti, C. Kruegel *Abusing Social Networks for Automated User Profiling, March 2010*
`http://iseclab.org/papers/socialabuse-TR.pdf`

[3] C. Sumner, *Social Networking Special Ops: Extending data visualization tools for faster pwnage*
`http://www.securityg33k.com/wp/BH10.pdf`

[4] L. Bilge, T. Strufe, D. Balzarotti, E. Kirda, *All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks, April 2009*
`http://www.iseclab.org/papers/www-socialnets.pdf`

[5] *Facebook developers get access to mobile phone and address information, January 2011* `http://www.eweek.com/c/a/Security/Facebook-Developers-Get-Access-to-Mobile-Phone-Address-Information-195408/`

[6] *Social networking websites reviews, 2011*
`http://social-networking-websites-review.toptenreviews.com/`

[7] Department for Communities and local Goverment,UK, *Online social networks, research report, October 2008*
`http://www.umic.pt/images/stories/publicacoes2/Online_Social_Networks.pdf`

[8] *Facebook statistics, January 2011*
`http://www.facebook.com/press/info.php?statistics`

[9] *Facebook description, January 2011*
http://en.wikipedia.org/wiki/Facebook#cite_note-500m-5

[10] *Facebook privacy policy, December 2010*
http://www.facebook.com/policy.php

[11] *Facebook Statement of Rights and Responsibilities, October 2010*
http://www.facebook.com/terms.php

[12] *Facebook Data Mining, January 2011*
http://en.wikipedia.org/wiki/Criticism_of_Facebook#Data_mining

[13] *Twitter User Statistics, September 2010*
http://twitter.com/about

[14] *Twitter Privacy Policy, November 2010*
http://twitter.com/privacy

[15] *Investment of Goldman Sachs Technologies on Facebook, January 2011* http://finance.fortune.cnn.com/2011/01/03/what-does-goldmans-investment-in-facebook-mean/

[16] *The value of a Facebook fan: an empirical review, June 2010*
http://www.syncapse.com/media/syncapse-value-of-a-facebook-fan.pdf

[17] *Facebook credit cards, January 2010*
http://www.usatoday.com/tech/news/2010-09-01-target01_ST_N.htm

[18] *Endomondo pro version landed, December 2010*
http://www.endomondo.com/blog/pro-version-landed

[19] *Social networking statistics, Lemieux and associates, 2010*
http://www.lemieuxassociates.com/Posted%20Documents/SIIAPresentationNotes.pdf

[20] *Social media statistics, Privacy and safety, 2008*
http://socialmediastatistics.wikidot.com/privacy-and-safety

[21] *Court finds that "private" posts to social network sites are not confidential and orders user to disclose log-in names and passwords, November 2010*
http://www.internetecommercelaw.com/articles/privacy/

[22] *Arab women being blackmailed over online photos, January 2011* http://www.thenational.ae/news/uae-news/arab-women-being-blackmailed-over-online-photos

[23] *Lenders using Facebook, Twitter to gather borrower information, May 2010*
http://www.post-gazette.com/pg/10148/1061287-28.stm

[24] *How can social networks affect your health insurance rates, 2011* http://www.reputationdefender.com/how_to/how-social-networking-can-affect-your-health-insurance-rates/

[25] *The Inside Story of How Facebook Responded to Tunisian Hacks, January 2011*
http://www.theatlantic.com/technology/archive/2011/01/the-inside-story-of-how-facebook-responded-to-tunisian-hacks/70044/

[26] *Fired Over Twitter: 13 Tweets That Got People CANNED, January 2011* http://www.huffingtonpost.com/2010/07/15/fired-over-twitter-tweets_n_645884.html#s112801&title=undefined

[27] *Man gets fined for insulting mayor via Twitter, July 2010* http://tweakers.net/nieuws/68835/man-krijgt-boete-voor-belediging-burgemeester-via-twitter.html

[28] *Social media and law enforcement; Who gets the data and when, January 2011* https://www.eff.org/deeplinks/2011/01/social-media-and-law-enforcement-who-gets-what

[29] *Google Crawl Page, January 2011* http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=34439

[30] *Web 2.0 Explained, January 2011* http://en.wikipedia.org/wiki/Web_2.0

[31] *Facebook Data Mining, January 2011*

[32] *EXIFtool, January 2011* http://www.sno.phy.queensu.ca/~phil/exiftool/

[33] *PyCurl Python Extension, January 2011* http://pycurl.sourceforge.net/

[34] *Scrapy Python Extension, January 2011* http://scrapy.org/

[35] *Vizster Research Paper, January 2011* http://vis.berkeley.edu/papers/vizster/

[36] *Current Vizster Development Page, January 2011* http://hci.stanford.edu/jheer/projects/vizster/

[37] *Wikileaks' Twitter account details sought* http://www.cbc.ca/world/story/2011/01/08/wikileaks-twitter-subpoena.html

[38] *How social networking can affect your health insurance rates* http://www.reputation.com/how_to/how-social-networking-can-affect-your-health-insurance-rates/

[39] *Facebook Opt-Out Example 1* http://tweakers.net/nieuws/72331/facebook-introduceert-places-deals-in-europa.html

[40] *Facebook Opt-Out Example 2* http://tweakers.net/nieuws/72039/facebook-apps-kunnen-meer-privedata-gaan-opvragen.html

[41] *Facebook Opt-Out Example 3* http://tweakers.net/nieuws/72277/facebook-verkoopt-userpostings-zonder-opt-out-te-bieden.html

[42] *EFF Law Enforcement Overview* https://www.eff.org/files/EFF_Social_Network_Law_Enforcement_Guides.xls

[43] *Social Network Blackmail 1* http://security.nl/artikel/36048/Man_chanteert_tiener_met_naaktfoto's_op_Facebook.html

[44] *Social Network Blackmail 2* http://security.nl/artikel/35821/1/Man_steelt_naaktfoto's_uit_gekraakte_e-mailaccounts.html