Universiteit van Amsterdam
System & Network Engineering

The devil is in the details: Social Engineering by means of Social Media

Research Project

Authors:
Yannick Scheelen
yannick.scheelen@os3.nl

Daan Wagenaar
daan.wagenaar@os3.nl

Supervisors:
Marc Smeets
msmeets@os3.nl

Marek Kuczyński
Marek.Kuczynski@os3.nl

July 11, 2012

Abstract

Corporations often spend a lot of money on technical security measures in order to secure their data. However, often overlooked are social engineering attacks. A difficult aspect of these type of attacks is that they cannot simply be prevented by implementing a technical solution. This is because they target humans instead of computer systems.

One of the most important aspects of social engineering is to have a solid base of information. This information is not only needed to find the right person to social engineer but also to find the right angle of approach, as well as to aid in the overall successfulness of the attack.

In this paper we research whether the aggregation of information, found on multiple social media networks, will lead to more successful social engineering attacks. We do this by targeting a specific company and aim to find as much information as possible on its employees.

We aim to prove that the combination of existing information gathering techniques for online social networks can be used to create a detailed set of information of the company and its employees. In detail, the research presented in this paper focusses on collecting and matching data found on LinkedIn and Facebook. This is done by employing a combination of zombie profiles and cloned profiles in order to crawl both public and private user data. Social synergy between both LinkedIn and Facebook is used to extend the amount of information found for a single person. To prove the detailed knowledge of the target company we create a hierarchical view of the company in which employee relations are depicted. In addition, detailed information per employee is displayed in this visualization for the purpose of improving social engineering attacks. We also research the dangers companies face when such detailed employee information is available as well as how to protect against these dangers.

We conclude that it is possible to automate the aggregation of data from different social media networks. Furthermore, we conclude that the obtained data is detailed enough to create a hierarchical view of a company. Finally we propose a combination of techniques and procedures to mitigate these dangers and attacks.

Contents

1 Introduction

1.1 Research Introduction

1.1.1 Social Engineering

Social Engineering (SE) is the art of manipulating a person in such a way that a target performs an action or reveals private information that he or she would otherwise not do. Lately, SE has gained increased attention in the hacking scene as more and more complex Information Technology (IT) security systems are deployed. Famous social engineer Kevin Mitnick explains in his book [1] that it is often easier to trick a person into revealing a password than to crack someone's password. This means that even when a company spends a lot of resources to get their IT security in check, they can still be vulnerable to SE. This exact scenario has been tested in [2], where a case study shows that SE can defeat a company's high-tech IT security measures. The techniques used in SE are diverse and extensive as explained in [3]. However, the basis for every SE attack and for every social engineer is information. As much information as possible on the target is collected and no fact or detail is unimportant. Having a solid base of information on the target can mean the difference between convincing someone to hand over information, or being denied of that information.

1.1.2 Online Social Networks

In 2010, Steven van Belleghem showed in [4] that 72% of all Internet users are part of at least one Online Social Network (OSN) site, which translates to 940 million users worldwide. We can therefore safely say that OSNs are a part of almost everybody's everyday life. OSNs like Facebook, MySpace, Twitter, Google+ and LinkedIn allow their users to share all kinds of information publicly or amongst friends and colleagues. The sharing of this information in combination with the widespread use of social media sites make it a wealth of information for social engineers, advertisement companies and other interested parties, all of them who can use this data to more specifically target persons.

Some differences between social media sites exist. For example, LinkedIn aims to provide a social media platform for a person's professional life. It allows for the users to share educational and work history as well as current employment. Facebook on the other hand tries to provide a social media platform for a person's personal life. Allowing for the sharing of personal information like photos, likes, relationships, etc. Twitter aims at the quick sharing of small status updates to the wide public whilst Google+ allows you to share personal information between different sets of people, called circles. Services like Foursquare are popular for their geotagging capabilities.

When people share their information it can effectively mean three things. First, the data shared is truly public, meaning that anyone that can browse the Internet can find the information, even if they don't have an account on that specific OSN. Second, the data is publicly available for people that also have an account on that specific OSN but are not connected to the person sharing the information. Finally, data can be shared in a private setting meaning that only the people that have an account on the same social media site and those whom are also connected to the person sharing their information, can view the information. This is often called a `friend` or a `connection`, in OSN terms.

An ever increasing problem that OSN faces is the use of fake accounts, also known as zombie

profiles. These fake accounts are often operated by people that aim to gather private information by tricking the user into accepting the friend requests send from these fake profiles. Once a user accepts such a friend request, the operator of the fake profile has access to information it did not have before. A similar approach is to 'clone' a user's profile of one OSN to a different OSN on which the user is not registered. By doing so, one can trick the friends of the original user's profile to be friends with the cloned profile on the second OSN. This approach and its effectiveness are based on the fact that the friends of the original user's profile are likely to accept the cloned profile as they will believe that the original user has just created a new account on the second OSN.

1.1.3 Importance of information

As previously explained, different OSNs have different goals and allow for the sharing of different kinds of information. This means that a social engineer could get a more complete set of information if it searches different OSNs for the same user. By mixing and matching the information found on different OSNs, a more complete set of information can be compiled that otherwise would not be possible.

With the possibility of gathering a substantial amount of information from these OSNs, a core concept of social engineering is touched. As said in [3], when it comes to the security of an organization's data and infrastructure, social engineering can already be done by only 'knowing the right people'.

1.2 Research questions

The focus of our research, as explained in the previous section, brings us to the following research question:

> *How can Online Social Networks be used in the automated creation of a graphical view of the company hierarchy and its employees for the purpose of social engineering?*

The subquestions listed below will help us to answer our main research question.

- How can current information gathering techniques be combined to achieve this goal?
- What are the consequences for companies?
- What can companies do to mitigate this process?

1.3 Research scope

Our research describes the processes needed for creating a crawler that is capable of crawling the employees of a specific company. The result is an overview of all the employees found as well as their current position and any other relevant information.

Next, a crawler capable of crawling Facebook profiles based on the information collected from the LinkedIn crawler will be created. The aggregation process of our program will be able to enrich the individual information of a specific user profile if a match should be found.

Given that a part of the research investigates the possibility of creating hierarchical business relations between employees by employing publicly available data, the social network site LinkedIn

has been chosen as a starting point. LinkedIn being the largest business-oriented OSN can provide us with the necessary information.

We chose Facebook specifically for the data enrichment aspect because it is, nowadays, the most active and largest social media network. We expect to retrieve more information from a social media network like Facebook, than from other, less used social networks such as Google+, Twitter or Foursquare.

We aim to find out whether the automation of graphing the company hierarchy and its employees is feasible. However, the actual automation of every technical action is not subject of our research. We therefore have chosen to perform some actions manually when it's already been proven that these specific actions can be automated.

1.4 Document structure

This research provides a structured and methodological approach in how OSNs can be used in the automated profiling of a company and its employees for the purpose of social engineering.

Before explaining our research methods, a theory section is presented which clearly explains the concepts that are going to be handled throughout the research paper. The theory section will describe the current privacy situation for both LinkedIn and Facebook, will distinguish between the different user groups that use these social networks and how they are used. Furthermore, it will explain an outline of the information gathering techniques that are going to be employed in the research methods section (section 3).

The research methods section details and explains how the necessary information for the research is obtained. It starts by explaining how the crawlers are created, with what specifics in mind, and how and which data is stored. Afterwards, the information gathering techniques described in the theory section are explained in practice through a set of flowcharts that depict how the combination of the crawlers and profile creation techniques can be best put to practice to gather a maximum amount of information.

The gathered information and retrieved results are explained in section 4, called 'Results of the research'. These results will be backed up in the 'Analysis of the research' section, section 5, where they will be explained in detail and in concurrence with the research methods.

Finally, a conclusion is given that presents our final findings and answers the research questions as posed in section 1.2.

2 Theory

Understanding how social media networks individually handle the privacy of their users is an important aspect for this research. To get a maximum amount of personalized data from the profiles that are going to be crawled, it has to be known how much information is visible either publicly or privately. Knowing the settings that define the way information can be hidden or unhidden on profiles, will allow for us to create a flow of actions that have to be taken in order to maximize the amount of data that can be retrieved. Sections 2.2 and 2.3 will provide the details for both LinkedIn and Facebook's privacy settings as they are today. Finally, we describe different building blocks that are going to be used throughout the research, which will be deployed to obtain the necessary information for our research.

2.1 Necessary data for Social Engineering

One of the most important aspects of Social Engineering is information. It has been said in [5] that no information is irrelevant, even the slightest detail can lead to a successful social engineering breach. In his book: 'Social Engineering: The Art of Human Hacking', Christopher Hadnagy explains how vital information is for an efficient social engineering attack. We build our research based on the premise that having more information and more detailed information on a target increases the chances of a successful social engineering attack.

Following a taxonomy of social networking data presented by Bruce Schneier[6], we define different types of user data that are important for our research and that we want to collect. We will focus on 'service data', 'disclosed data', 'entrusted data' and 'incidental data'. These types of data can be found with a single snapshot of a user's profile page. Schneier further defines 'behavioral data' and 'derived data' as possible data types, this data has to be interpreted and needs multiple recordings before it can be used.

- **Service data**: Data the user provides to the OSN in order to use it (name, date of birth, e-mail address)
- **Disclosed data**: Information the user posts on his/her own pages
- **Entrusted data**: Information the user posts on other peoples' pages
- **Incidental data**: Information that is posted on the user's pages by others
- Behavioral data: Data the site collects about the user's habits by recording what he does and who he does it with
- Derived data: Data that is derived from all the other data

Not all of this information is publicly available however, tables 2 and 3 give a full overview of how the privacy settings are defined for the OSNs we will be using. Getting to information that is not publicly available requires different techniques as explained in section 2.4.

2.2 LinkedIn

2.2.1 General information on LinkedIn

LinkedIn is a social network which is mainly used for professional networking. As of March 31, 2012, LinkedIn operates the worlds largest professional network on the Internet with 161 million members[7]. According to the website's published statistics LinkedIn is being very actively used, stating that LinkedIn members did nearly 4.2 billion professionally-oriented searches on the platform in 2011 and are on pace to surpass 5.3 billion in 2012. According to Alexa, LinkedIn is currently ranked as the 12th most visited website in the world[8].

2.2.2 LinkedIn's user relations structure

LinkedIn has created a unique scheme to identify how well a user knows other users. It defines the relations between its users using a tier-based system divided in three tiers. The three tiers are designed as such, that a user is either:

1. 1^{st} tier: directly connected to another user (1-1 relation)
2. 2^{nd} tier: connected through the 1^{st} tier
3. 3^{rd} tier: connected through the 2^{nd} tier
4. Out of Network: not connected

Despite the three-tier approach, LinkedIn remains a scale-free network. The concept of a scale-free network has been discussed several decades ago[9] but has only recently been applied to modern day networks, when Albert-László Barabási mapped the topology of a portion of the World Wide Web[10]. A scale-free network has a degree distribution that follows a power law as proven in [11].

The importance of the scale-free property of LinkedIn's network lies in the fact that the amount of connections one user has, expands with a power law distribution relative to the amount of connections his first and second tier connections have. As figure 1 illustrates, a user's network expands rapidly when new connections are made, which results in a larger set of data to be available for the user.

One of the ways LinkedIn wants to protect the privacy of its users, is by restricting information to be viewed by people that are Out of Network. Generally, this is done by obfuscating the name of the user when general searches are made. For example, a company-wide search will only yield a 'LinkedIn Member' result, instead of the real name of the LinkedIn member. Only when another user becomes connected to the employee of the company through at least a 3rd tier connection, will the name obfuscation be gone. Even then, on a basic account, LinkedIn only allows you to view full names of 1^{st} and 2^{nd} degree connections. Third tier connections will appear as first name and the first letter of the surname. Table 1 provides a full overview of which aspects of a LinkedIn profile that can be viewed per tier.

As a side note, LinkedIn offers several account type upgrades which can be bought and provide an extended amount of details that are viewable when searching profiles and companies. The findings that are presented here come from the perspective of a basic account, showing which information a normal user can obtain by using LinkedIn.

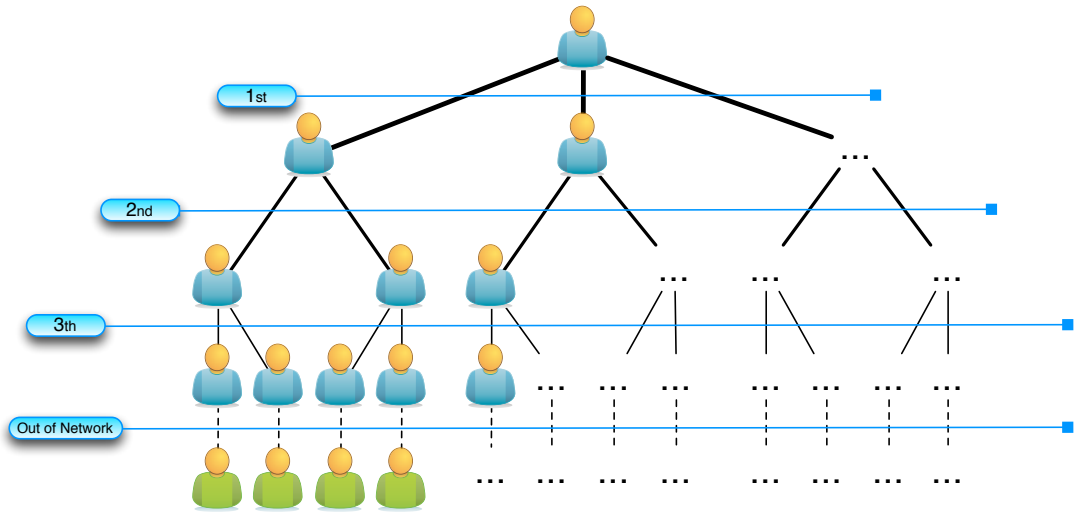


Figure 1: A graphical representation of LinkedIn's three-tier connection scheme

	1st tier	2nd tier	3rd tier	OON	Groups
Name	√	√	√	√	√
Headline	√	√	√	√	√
Summary	√	√	√	√	√
Location & Service	√	√	√	√	√
Current position	√	√	√	√	√
Education	√	√	√	√	√
Connections	√	Only amount	Only amount	Only amount	Only amount
Public profile link	√	√	√	√	√
Experience (Full)	√	√	√	-	√
Languages	√	√	√	-	√
Education	√	√	√ (Limited)	-	√
Additional Information	√	√	√ (Limited)	-	√
Skills	√	√	√	-	√
Summary	√	√	√	-	√
Contact for given reasons	√	√	√	√	√
Recommendations	√	-	-	-	√
Contact information	√	-	-	-	-
Who you can directly connect to	n/a	√	-	-	√

Table 1: Overview of the information each tier and group members can view of a default LinkedIn profile

2.2.3 Connecting to users on LinkedIn

Following the theory of the scale-free network, having a new account on LinkedIn without any connections means that the user will not be connected to anyone. It is only by creating connections to other users that the personal network will expand and more information on other users will become available. When information on a company is needed, it will not suffice to just create a zombie-like account on the network. The account has to be able to connect to a company's employee to expand the network sufficiently to get unobfuscated results.

Creating connections is done by sending connection invitations to other users of the network. LinkedIn imposes several restrictions when it comes to sending invitations, starting by the fact that invitations can only be sent once to a specific user. If the receiving user chooses to ignore the initial connection invitation, the requesting user would have to provide the e-mail address of the user he wanted to connect with in order to send a second invitation. If the receiving user accepts the invitation the connection is instantly made. Another restriction that LinkedIn has implemented, is by making sure that once a user is completely out of network, it becomes impossible to send a connection request to that user unless the full name is known. LinkedIn is still a social network and it has to provide the possibility of connecting with people that you know, but that are not yet in your professional network. By providing the full name of the user you'd want to connect with, it becomes possible to send a connection invitation. When making generic searches for companies and suchlike it will depend on the tier level whether or not another user can be invited to connect, as detailed in table 1. The only possibility of contacting a user that is out of network and whose name is not fully known, would be through LinkedIn's built-in messaging system: InMail. Unfortunately, sending InMail is a privilege reserved only for users who've upgraded their account to anything different than the basic package.

On LinkedIn, users can follow a company to keep in touch with any status updates the company goes through; updates being sent out by the company, profile updates of current employees at that company, or new arrivals at the company. A user automatically 'follows' the company that he sets on his profile. If you say you work at a specific company, you will automatically follow this company and show up in the company's 'Followers' list.

2.2.4 Public profile possibilities

LinkedIn has the concept of public and private profiles. The public profile is the profile that is crawlable by search engine spiders (Google, Yahoo, Bing,..), resulting in the fact that the information that is published on the public profile will be readable to everyone. The public profile's settings allow you to show specific parts of your profile and can be either disabled completely, or enabled. When it is enabled, the basic information of your profile is always shown (Name, headline, industry, location, number of recommendations) and the user is then able to specify which parts of the public profile he wants to publish. The parts that are publishable conform to the main sections in table 2, with an additional option of showing minor details. By default, the public profile is enabled and shows *all* the information the user has set on his profile.

2.2.5 Possible privacy controls

Next to the public profile, a user has the possibility to customize his own profile. The information that is shown on the public profile will be taken from the information that is set on the user's

private profile. Table 2 provides a full overview of the default privacy settings that LinkedIn sets on user profiles and the control a user has on which fields he wants to fill in and to what extent. Basically, the only fields that are compulsory are the 'Display Name' and the Headline'. Every other field is optional and the amount of detail that can be provided depends on the user. Even though the user has control over which fields he fills in, not every field has its own privacy setting. Moreover, most of the general sections can be either shown or not, with no refinement for specific privacy controls.

2.2.6 Using LinkedIn for companies

In 2008, LinkedIn announced 'Company Profiles'[12], a feature that transforms every company on LinkedIn to having its own, independent and customizable profile page. Only current employees are eligible to create a Company Profile for the company they work for and in order to create a Company Profile, you would need to provide a registered, company-owned email domain.

Registered users can freely add a current job position and employer to their own personal profile. When a Company Profile is made, it is automatically indexed by LinkedIn's search engine. This means that when a Company Profile exists and a user wants to add a specific, existing company to its profile, he will get a list of search results to choose from. When choosing a company whose Company Profile already exists in the database, the user automatically gets added to the Company Profile as a currently working employee. Through this feature, companies have the possibility of closely following up on its current employees.

2.2.7 A company's default privacy settings

Company Profiles are profiles that list a company with all its attributes and values as set by the administrator of the Company Profile. When creating a Company Profile, a certain amount of settings are mandatory to be filled in (i.e.: Company size, location, main industry, description, etc...). These settings are all public information and can be viewed by any LinkedIn user. Any additional attributes that are set will become public information, with no options for setting privacy controls on the fields.

Finally, if you do not have a registered company owned e-mail address to create a Company Profile you may still create a group to promote your business. Groups are mainly formed based on interests and tries to connect users who share an interest or a passion, supported by a limited form of a discussion area. However, companies are using Groups as well to try and connect all their employees and everyone who has an interest in the company. Group pages can be either open or closed, with closed being the default setting. Group managers can allow or deny access to the Group for users that request access. Only users that are registered with the Group (the Group being either open or closed) can follow and participate in discussion that are being held on the Group's page. When users share a Group they will be able to view more information on the profiles from one another, as table 2 lines out. Important to note is that if a Group connection is 3^{rd} tier or Out Of Network, the last name will still be partly obfuscated.

Item	Default	Options		
Basic Information				
Display Name	**Full name**	First name, Last name first letter		
Headline	**Show to everyone** (user customizable)			
Show connections	**Your connections**	Only you		
Activity Broadcasts	**Your connections**	Everyone	Your Network	Only You
What others see when viewing their profile	**Name and headline**	Anonymous profile characteristics		Totally Anonymous
Public profile	**On**	Off		
Accept Introductions and InMail	**Introductions and InMail**	Only Introduction		
Profile Photo	**Everyone**	My Connections		My Network
Personal Information				
Birthday	**My Network**	My Connections		Everyone
Birth year	**My connections**	My Network		Everyone
Marital Status	**My connections**	My Network		Everyone
Other information	**Show everything**	None		
Education				
All	**Show everything**	None		
Experience				
All	**Show everything**	None		
Summary				
All	**Show everything**	None		
Languages				
All	**Show everything**	None		
Additional information				
All	**Show everything**	None		
Skills & Expertise				
All	**Show everything**	None		

Table 2: Overview of the default LinkedIn privacy settings, items in bold indicate the default setting that is used by LinkedIn with a new account. The other items shows the possible privacy options that can be set per item.

2.3 Facebook

2.3.1 General information on Facebook

Facebook is the world's largest Online Social Network, with 901 million active users at the end of March 2012 and counting. Not only does it have the largest user base, Facebook reports that more than half of its registered users avidly use the networking site, with 526 million daily active users on average in March 2012[13]. Similarly to LinkedIn's network, the Facebook's network can be described as a scale-free network, which translates to more than 125 billion friend connections at the end of March 2012. According to Alexa, the website currently ranks second among all websites in the world[14].

Contrary to LinkedIn's professional focus and business-oriented set-up, Facebook focusses primarily on the individual and does so by providing extensive features that will make every single

profile as personal as possible. It provides input fields that cover almost every aspect of a user's personal life, input fields that can optionally be expanded to the user's will. A full overview of every detail that can be entered by a Facebook user is given in table 3 and the only mandatory information that has to be given is the user's name, gender and birthday. Facebook limits its audience to users of age 13 and above, which is why the date of birth is of importance.

2.3.2 Connecting to users on Facebook

Connecting to people on Facebook works by 'friending' other users on the social network. Differently than LinkedIn's three-tier system, there are no restrictions set on who you can send the so-called 'friend requests' to. By default, everyone on Facebook can receive friend request from every other user. There is, however, a possibility of allowing only 'Friends of friends' to send you friend requests. The 'Friends of friends' notion is a way Facebook wants to restrict information and features to only the people inside a user's personal network. While it is a way to give the user a feeling of privacy, it allows for elaborate crawlers to obtain personal information by friending a friend of the user that is targeted. Without the target knowing, his or her information can still silently be collected.

2.3.3 Facebook's features and privacy settings

In September 2011, Facebook introduced 'Timeline'. Facebook's Timeline is the new profile page design which changes the old profile page from a list of your most recent updates, to a complete overview of all the photos, videos, and posts that a user is linked with, categorized according to the period of time in which they were uploaded or created. A list of years is displayed on the side of every Timeline, acting as an actual timeline from the day a user was born until now, listing every single event that has occurred over time in a chronological order.

The Facebook Timeline provides an incredible amount of user information, as table 3 shows, 71% of all the information that a user has filled in on Facebook will by default be visible for the entire Internet. In a graphical overview, Matt McKeon has illustrated that Facebook has frequently changed the default privacy settings over the years[15], supporting mounting criticism about the business policies of Facebook. This translated in officials advising the European Commission sending a letter to the social networking company, saying that changes to its default settings made in December 2009 were 'unacceptable'[16]. Facebook responded to and agreed with the criticism that were addressed to them, by making the privacy settings more transparent.

While LinkedIn doesn't allow its users to change the privacy settings of every input field but restricts it to some (table 2), Facebook does provide this possibility. In practice, this means that a user on Facebook now has complete control over what information he wants to share with whom. Yet, even though the privacy controls have become more transparent and accessible, there is still an difference noticeable over the years concerning the default privacy settings that are applied on the user's information, namely that Facebook is making more and more information publicly visible by default[15].

Referring back to table 3, the different types of user relations are laid out by proximity to the user. Information can be shared either completely privately (only you), to friends, to friends of friends, to all the Facebook users, or to the entire Internet.

Table 3: Facebook's default privacy settings, overview of which information which part of the network can see when you create a new Facebook account

Item	Friends	Friends of Friends	All Facebook users	The Internet
Basic Information				
Name	√	√	√	√
Picture	√	√	√	√
Current location	√	√	√	√
Hometown	√	√	√	√
Gender	√	√	√	√
Birthday	√	√	-	-
Interested in	√	√	√	√
Languages	√	√	√	√
About me	√	√	√	√
Contact information				
Emails	√	-	-	-
Mobile phones	√	-	-	-
Other Phones	√	-	-	-
IM screen name(s)	√	-	-	-
Address	√	-	-	-
Town/City	√	-	-	-
Zip	√	-	-	-
Neighbourhood	√	-	-	-
Website	√	√	√	√
Friends and Family				
Relationship Status	√	√	√	√
Family	√	√	√	√
Friends list	√	√	√	√
Education and Work				
Employer	√	√	√	√
University	√	√	√	√
Secondary School	√	√	√	√
Philosophy				
Religion	√	√	-	-
Political views	√	√	-	-
People who inspire you	√	√	-	-
Favourite Quotations	√	√	-	-
Arts and entertainment				
Music	√	√	√	√
Books	√	√	√	√
Movies	√	√	√	√
Television	√	√	√	√
Games	√	√	√	√
Sports				
Favourite sports	√	√	√	√
Favourite teams	√	√	√	√
Favourite athletes	√	√	√	√
Activities and interests				
Activities	√	√	√	√
Interests	√	√	√	√
Likes	√	√	√	√
Other				
Wall Posts	√	√	√	√
Photos	√	√	√	√
Who can send you messages	√	√	√	√
Who can send friend requests	√	√	√	√

2.4 Data collection techniques: Building blocks

The following subsection describes the different approaches that can be taken to create profiles on social networks. There are different approaches and trains of thought and each of them has a specific value and goal. There is a correlation found between matching information on different social networks and cloning profiles. Using the information on one network and transposing it to another is explained in subsection 2.4.2 and can also be applied to the concept of Reverse Social Engineering, as explained in subsection 2.4.3. A different method that is explained handles the creation of random fake profiles, with the intent to connect to a maximum number of users and gather as much information as possible by infiltrating the user's private network.

2.4.1 Synergy of social networks

Matching the information that is available on different online social networks is a concept that has been worked out in the work by Kuczyński et al[17]. Through matching similar identifiers in different social networks, it becomes possible to create an enhanced profile of a certain individual.

2.4.2 iCloner

The idea of using fake profiles for infiltrating online social networks has been touched in a different context by Bilge et al[18], where a user profile is cloned based on the extracted information of its account on one social media network. With that gathered information their program, iCloner, would duplicate the account on a different social media network site and create friend connections to people also found on the first network.

2.4.3 Social Engineering

The classic use of social engineering in social media networks has been described as a data collection technique by Bonneau et al[19], where 'Targeted Compromise', also known as 'spear-phishing' attacks is proven to be a highly effective technique. With a targeted attack on a subset of a social network site, 50% of the user profiles of the subset was fully viewable with an amount of 0.16% of profiles, where the full subset is 100%. This means that less than 1% of targeted profiles has to be made in order to already retrieve a high volume of information from a social network site. Furthermore, the exploiting of the 'Friends of friends' feature on social network sites such as Facebook, allows for a network discovery that is around 100 times faster.

Another aspect to using these fake profiles efficiently, is the fact that they should be built up intelligently based on a target's interest, hobbies and other personal information, in a way so it would look trustworthy and real to the target. Using the techniques described by Irani et al.[20], Reverse Social Engineering (RSE) could be used to trick the target with the new and appealing generated profile. Tricking the user would involve making a fake profile in such a way that it would show in the 'People you may know' section of Facebook. The term 'Reverse' social engineering is coined because it requires the target to be so enticed in the fake profile that he would send the friend request himself, instead of having the profile creator do it.

2.4.4 Zombie profiles

A zombie profile in the context of Online Social Networks is a fake profile that is created on the network. It is filled with purely random information and has the goal to create as many connections as possible with other users on the same network. If a connection can be made, the employer of the zombie profile obtains access to the private information of the victim. In a research conducted in 2011[21], an automatic approach to creating a botnet of zombie profiles has been invented, named Socialbot. The researchers have shown how zombie profiles can be created, deployed and commissioned automatically on a large scale.

2.5 Summary

In this section we have defined which information we want to gather and how much of this information is available by default on LinkedIn and Facebook. By looking at the privacy settings these networks employ on the default profiles, we have specified which information will be easy to gather and which information will require further actions before it becomes available.

We determined different building blocks that are necessary to retrieve every bit of information we need to finally compose a hierarchical visualization of a company and its employees.

In the following section the implementation of these building blocks is going to be explained.

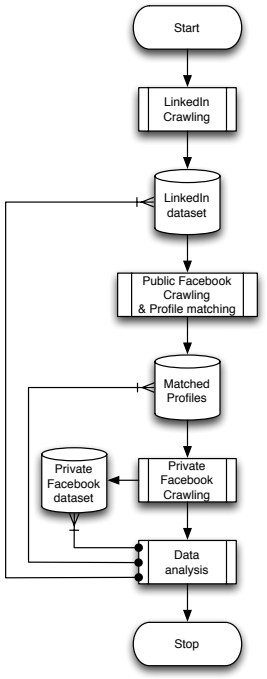
3 Research Methods

This section explains the methods and approaches we use for gathering our data. We start by giving a general overview of our approach, followed by an in-depth view on some of the more technical aspects regarding the information gathering processes. During this section we define `targeted company` as the company and its employees being targeted, and `friend(ing)` as the process of sending a 'friend' request to a user. Lastly, we incorporate the ways we investigate the risks and mitigations of social engineering by means of interviews and related research on the topic.

3.1 General approach on the research

We have subdivided our experiments into several different sections. By doing so we allow for a more modular approach in which we can fine tune each aspect of our research approach. The entire process can be seen in figure 2, in which four different processes are shown, namely; LinkedIn crawling, Public Facebook crawling & Profile matching, private Facebook crawling and finally the data analysis. This flowchart shows that three of these processes produce datasets and that two of the processes rely on these datasets. Furthermore, the data analysis process requires the data found in all three datasets in order for the analysis to be able to take place.

Figure 2: The overall experiment process



Each of the defined processes can be further subdivided into more complex flows to better illustrate the process of our experiments. Figure 12 in appendix A shows the LinkedIn crawling process in more detail. The first two steps in this process involve the creation of a zombie profile on LinkedIn with authentic looking information. Important in these two steps is that the profile must appear as if it is working for the targeted company. All other details just need to appear believable to convince people that the profile is genuine. Using this account we can provide the name of the targeted company and search for a company profile and its followers. As explained in section 2, users on LinkedIn can follow companies. The list of followers for our target company will be targeted and the profile pages of the followers will be parsed and stored. Finally a connection request will be sent to each of the parsed followers. Once one or more connection requests have been accepted, we can continue with searching for people who work for the target company by means of the default search function of LinkedIn. From these search results, all the 2^{nd} tier connections will be parsed, stored and invited to become friends. We can then once more check whether our invites have been accepted or not and if so redo this process to get more 2^{nd} tier information and friends. Once enough results have been gathered, we can stop this process.

With the initial LinkedIn dataset we can move forward onto crawling Facebook for public user information, this process can be seen in figure 13 in appendix A. For each of the persons found in the LinkedIn dataset a separate search on Facebook is performed. The results returned from the search will be processed and matched against the information found in the LinkedIn dataset. We try to determine whether a Facebook profile matches the record in that dataset based on the publicly available data. We do this based on the fields described in previous work, that stated that a match between name, profile picture and location can be made. We will extend these fields to also incorporate the current work place or past education among others.

3.2 Matching profile information on LinkedIn and Facebook

3.2.1 Matching using public data

The first step will be to try and match profiles from LinkedIn to profiles on Facebook by using only publicly available data. We will look at the information an employee posts on his LinkedIn profile and see if there are users on Facebook who have identical information and have this publicly displayed. We will do the initial search on Facebook based on the employee's name as taken from LinkedIn and then see if there are any additional fields that match. If a match can be made, these profiles will be identified as 'matched'.

3.2.2 Matching using FLEMP

If initially we are unable to use publicly available data to make a match, different actions can be taken to try and make a match in a different way. As figure 13 shows, an iteration happens on the friend lists of the profiles we have already matched.

We devised a technique called 'Friend List of Earlier Matched Profiles' (FLEMP), which will search through the publicly available friend lists of the profiles that have already been publicly identified. This technique compares names found in these friend lists to the names of unidentified profiles in our dataset. Our hypothesis is that when employees use Facebook there is a chance that they also friend their colleagues. More precisely, employees are likely to have some of their colleagues on Facebook with whom they are also connected on LinkedIn. Thus, if we are able to match a LinkedIn profile to a Facebook profile based on its occurrence in a friend lists, the profile will be put on the list of Facebook matches.

In addition, FLEMP can also be used on the friend lists of people that are 'friends', this because friend lists of friends are most of the time available (as can be seen in table 3). However, during this research FLEMP is only used in combination with publicly available friend lists.

3.2.3 Matching using Zombie Profiles

The third and final technique that is used to try and match a Facebook profile to a LinkedIn profile, is to make use of zombie profiles. We manually create and populate zombie profiles where needed. We have chosen to do this manually because it is already proven that it can be automated on a large scale in a recently published paper[21]. To create a believable name, we use an online service called 'naamgenerator.com'[1]. This website also generates a valid home

[1] www.naamgenerator.com

address, phone number and e-mail address that we can use as content for the zombie profile. On LinkedIn, any additional content has been filled in with valid random data that applies to the context of the function of the targeted company. For example, when targeting a financial company, the education field could be filled in with 'Master in Finance and Investment'. Finally, each zombie profile will be of the male gender as results collected in [21] prove that users are less likely to report a zombie profile as spam, when the gender is male.

It often happens that a user's name on LinkedIn has multiple occurrences on Facebook. To try and find the right profile that matches with the LinkedIn profile, Facebook zombie profiles filled with completely random information, will send friend requests to all the remaining and unmatched found results. If a friend connection is made, we will have the ability of accessing the private data of that profile. With that information we can try again to match the Facebook profile to the LinkedIn profile. If this method does not yield in any matches either, we deem the LinkedIn profile unmatchable and continue with a different profile.

3.3 Information Gathering

This section will discuss the different aspects and specifics of how we gather results on LinkedIn and Facebook. Furthermore, it shows which specific fields we have crawled for both LinkedIn and Facebook.

3.3.1 Crawling technology

For the crawling specific actions we make use of Scrapy[2]. Scrapy is a Python-based framework for scraping and crawling webpages by using so called XML Path Language (XPath) selectors. An XPath selector is a specific way of identifying the different elements, tags and nodes inside a HyperText Markup Language (HTML) or eXtensible Markup Language (XML) document. The reason for choosing Scrapy to support us in our crawling needs, is because of the fact that it is a well-maintained and updated product and that the use of Python is preferred since knowledge of this language is present within the project group. All the data that is going to be gathered during the research will be saved in a MySQL database.

3.3.2 Data obfuscation

Because of the privacy sensitive nature of the crawled information, the actual data that is stored will first be undergoing a data obfuscation technique. All data will be put through a hashing function after which it will be human unreadable. The properties of the hashing function allows us to identify unique data. When the data is the same, the same output will be generated by the hashing function. This effectively means that we will still be able to see, for example, how many people have gone to the same school but we will not be able to see which school they went to. After the research has been concluded, the collected data will be integrally deleted.

[2] http://scrapy.org

3.3.3 LinkedIn search queries & results

As mentioned in the previous section, we will be performing searches on LinkedIn to gather as much information on the employees of the targeted company as possible. The search function provided by LinkedIn allows for a combination of search parameters and filters to fine tune the search results. This allows us to tailor the search results to the location of the targeted company. The fields used and their values can be seen in table 4.

Parameter	Value
Key words	Variable
Company	Company name
Postal code	Postal code of company location
Current and/or Past employment	Current employment
Distance	50 Miles

Table 4: LinkedIn search parameters

One important limitation is in effect while using the LinkedIn search function. A basic (free) LinkedIn account is only allowed to view 100 search results, effectively limiting the amount of information we can crawl per search query. To circumvent this limitation, we have used a number of different combinations of keywords to change the returned search results. However, this technique does not cause for every search query to return 100 unique results, in some cases the results can overlap and thus the results that are returned diminish as you try more and more different keyword combinations.

For us to focus on a specific location we fill in the postal code of the location of the targeted company and use a radius of 50 miles. This effectively means that the search results only show the profiles of people living within a 50 mile (80 kilometers) radius of the company location. We have chosen a 50 mile radius as a mobility research by the Ministry of Traffic[22] shows that the average distance an employee travels to his work in the Netherlands is 22 kilometers (14 miles). However, this number is an average and thus we want to use a larger value making sure we get as many relevant search results as possible.

3.3.4 Gathering data on LinkedIn

Before employees can be crawled, an entry point is needed to create the preliminary connection with the network. LinkedIn obfuscates the employee's name in company search results if the connection is completely outside of the network of the user who's making the search. The obfuscation comes in the form that the user names are replaced by the text 'LinkedIn Member'. In this case, it becomes hard to find all the affiliating users with the company. Another problem is the fact that once the users are completely out of network, it becomes impossible to send connection requests to those users.

To get around this issue, we found a small hole in LinkedIn's obfuscation, located in the 'Followers' page of a certain company, which is described in section 2.2.2. While a usual search for a company will provide an obfuscated list of employees, the 'Followers' list that LinkedIn provides removes part of the obfuscation. A 'feature' that LinkedIn provides is that a user automatically follows the company that he sets on his profile. If you say you work at a specific company, you will automatically follow this company and show up in the 'Followers' list. There is still a partial

obfuscation in place, yet now a first name and the first letter of the surname is given, together with the headline that usually indicates the current position held by the user.

By looking at this list of followers, we are able to get at least a part of the people who work at that company. Even better, the name in the 'Followers' list is a hyperlink that links directly to the user's public profile. This profile will show the full name, and all the details that are set to be seen publicly by the user, including the connect button that is normally replaced with an InMail button for users out of network. The logic behind this is that LinkedIn is still a social network and it will allow you to connect to people who are completely out of your network. The 'Followers' list follows the same logic, even though an exact name match hasn't been specified to query the followers list.

The fields we collect from Linkedin can be seen in table 5. We have selected these fields because these are almost always present on every profile whether you view them publicly or privately as can be seen in table 2. The crawling code is flexible enough to be expanded with additional fields easily if LinkedIn would introduce new fields or when other fields are deemed more important for social engineering.

LinkedIn Fields	
First name	Past employment
Last name	Education
User id	Websites
Headline	Twitter
Living location	Summary
Industry	Interests
Current employment	

Table 5: Crawled fields from LinkedIn profiles

3.3.5 Gathering data on Facebook

For each of the results in the subset of employees found on LinkedIn a separate Facebook search will be done. Depending on the results found for each LinkedIn profile specific actions will be taken. The flowchart shown in figure 13 in appendix A shows which specific steps and actions need to be taken in order to identify a match between a profile found on LinkedIn and a profile found on Facebook.

Only a select number of fields are used during the matching process. Kuczynski et al. discusses in [17] the use of the 'Name', 'Date of Birth' and 'Current City' in the matching process. We have extended this and also use the fields 'Current employment', 'Education', 'Past education' and 'Likes' in the matching process.

We define a successful match when two or more of the earlier mentioned fields match. However, we have defined three exceptions where only one field needs to match for a successfully identified profile.

Exception 1: If the 'Current Employment' field is the only field that matches.

Exception 2: If exactly one search result is retrieved and only one field matches.

Exception 3: If the profile pictures on both LinkedIn and Facebook are exactly the same.

Each found match will be stored and its public data will by crawled by our created Facebook crawler.

For Facebook we have decided to collect the fields as found in table 6. Once again, the addition of fields present on a person's profile is just a matter of adding the right code to the crawler.

Facebook Fields		
First name	Children	University
Last name	Relationship	College
User id	Interest in	High school
Profile link	Languages	Degree
Email	Activities	Company
Phones	Interests	Company Position
Gender	Music	Duration of employment
Birthday	Movies	Employment date
Current City	Sports	Employment description
Home town	TV Programs	Political view
Uncle & Aunt	Bio	Religion
Siblings	Quotes	
Wall	Friends	

Table 6: Crawled fields from Facebook profiles

3.3.6 Zombie Profiles for information gathering

For all the LinkedIn profiles that could not be matched but which did return results, two more methods will be employed to try and find a match. One technique is a so-called 'shotgun approach' meaning that we employ zombie profiles to spam friendship requests to all found results and wait for them to accept our requests. When these persons accept our requests we can crawl their private data.

3.3.7 iCloner on Facebook

The second technique involves sending friend requests to the matched profiles once again, but with one difference: we try and clone a profile found on LinkedIn, onto Facebook, using a technique described in the work by Bilge et al.[18].

The flowchart displayed in figure 14 in appendix A shows that an iCloned account will be created during the process of crawling private data from Facebook. We determine that a possible LinkedIn connection is iClone worthy, based on whether searching for that name on Facebook yields zero results. Of course, one could hide his Facebook profile from the search results but this is a risk we have to accept because you can never be sure 100% that a profile does not exist. We manually copy all the information from the LinkedIn profile onto the newly created Facebook profile. We do this manually because automating this process has already been proven successful in the previously mentioned paper. By doing so, we create a Facebook account that will be accepted more easily by the connections he or she has on LinkedIn, assuming these connections know the person sending the friendship request. If the friendship requests is accepted, we can crawl the private data of these profiles.

If we are unable to create a cloned account, we will use the same zombie profile technique mentioned earlier in this section. Note that we still store all the publicly available data for profiles that are matched on order to have at least some personal information of the user.

3.4 Company structure visualization

When all the data has been collected, a graph will be created that shows an interpreted functional structure of the company and which persons are connected to whom and by what function. This will be done by using a weight-based system in which we assign weights to the different functions that we encounter. We then create connections from the most-weighted function to the next most-weighted one and so on until we have no more functions to link. We then create links between the employees and their functions for a final representation of the company structure.

For the visualization of the company structure we will make use of a tool called Viszter [3]. This tool can, based on an XML file, plot a graph of how nodes are connected to each other. The XML file will be generated automatically from our database using a custom tool we will create. We have modified Vizster in such a way that visualizing the structure of the company and the aggregated information of the employees becomes an effective technique. By parsing the 'Current employment' field, we are able to get the job title of the employees and the department he or she works in. How this is done can be seen in figure 3, in which we show that three fields are reoccurring.

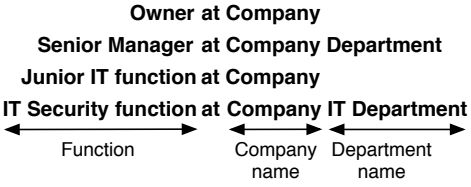


Figure 3: Parsing the current employment field

After the 'Current employment' field has been parsed, we assign numerical weights to each function based on the importance and hierarchical business placement of the function. To determine the weight each function receives, we apply the theory discussed in [23] along side common sense. The latter meaning, for example, that a Junior position gets a lower ranking than a Senior position.

By using these weighted functions we can define the connections between these functions and plot them in Vizster. After this has been done, connections will be made between the persons and their respective functions in their respective departments.

3.5 Interviews and literary research

Besides the technical aspects presented in the earlier sections, we will also be conducting interviews with different specialists active in the field of IT Security. All the specialists that we will be interviewing have an extensive amount of experience in social engineering. We conduct

[3] http://hci.stanford.edu/jheer/projects/vizster/

these interviews to try and gain insight into the benefits of our research for social engineering. In addition, we hope to gain insight as to what dangers our research can pose for companies. Furthermore, we hope to obtain information that could help us in understanding what companies can do to mitigate the processes researched.

Besides the interviews, a literature study will be done to hopefully strengthen the findings from the interviews. We hope to find concrete examples of social engineering attacks that could be made possible or improved based on the information aggregated by our research.

The combination of the interviews and the literary research will provide us with the necessary material to answer the biggest parts of our subquestions.

4 Results of the research

The summation of the results of our experiments are outlined in the upcoming section. They provide an analytical overview of all the results that were measured during the tests and provide the necessary knowledge to understand how the final analysis in section 5 is done.

4.1 Information on the targeted company

We conducted our experiments by targeting a company in order to get relevant results. We targeted a large corporation of about 5000 employees in total (in The Netherlands) but limited our scope to a specific branch of the organization. The targeted branch has about 300 employees spread across the country.

The total time spent on crawling data of this company was limited to two days. One day for crawling LinkedIn data and one day for crawling Facebook data.

4.2 Creating LinkedIn profiles and getting connected

4.2.1 LinkedIn zombie profile's connections

With our fake LinkedIn profile we sent out a total of 106 invitation requests.

Of the 106 connection invitations we sent out, 39 were accepted, resulting in a success rate of 36.7%.

Having 39 1^{st} tier connections results in our profile being connected to a total of 11,420 profiles (figure 4) according to LinkedIn. These are all profiles we would potentially be able to directly connect to.

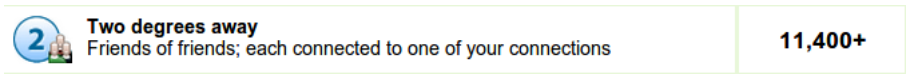


Figure 4: Total amount of profiles we were connected to

4.2.2 LinkedIn crawling results and Facebook matching

Our first filtering (based on keywords related to the targeted branch) resulted in 286 profiles that conformed to our search criteria, the information of all these profiles was crawled and stored.

Out of these 286 profiles, 125 LinkedIn users were found that have an account on Facebook, a total of 43%.

After applying a final filter on the dataset of 286 profiles a final dataset of 85 profiles was found. From this final dataset we found a total of 9 links to another OSN (Twitter).

The number of times we have crawled each field from a Linkedin profile can be seen in table 12 in appendix B.

Within this final dataset of 85 profiles, 37 users were matched on Facebook. This still conforms to the 43% that was found on the larger dataset.

The final search for Facebook matches resulted in another 9 Facebook profiles found, leading to 46 Facebook profiles out of 85 LinkedIn profiles, a total of 55%. For this final set of 46 Facebook profiles we have populated table 7 with the fields (and their combination) responsible for the matches.

Fields	Times matched
Current Employment	1
Current Employment, Education	13
Current Employment, Single result found	1
Current Employment, Education, Living location	9
Education, Living location	4
Education, Past education	1
Education, Living location	1
Found in Friend List of Earlier Matched Profiles (FLEMP)	6
FLEMP, Current Employment, Education	2
FLEMP, Living Location	1
Likes, Living location	1
Past, education, Living location	1
Exact profile picture	5

Table 7: LinkedIn to Facebook profile matching

Table 8 respectively shows the amount of profiles matched based on the different techniques used.

Match based on	Times matched
Public profile data	36
FLEMP	9
Zombie profiles	0

Table 8: Number of matches made per type

4.3 Creating and crawling Facebook profiles

4.3.1 Facebook iCloner

One iCloner account was used to connect to 10 people. Of these 10 friendship requests, 6 were accepted resulting in a total of 60%.

Furthermore, 4 Facebook users that work in the same company as our iClone profile sent a friendship request to the iClone profile. Of these 4 users, 3 of them were 2^{nd} tier connections to our LinkedIn zombie profile.

4.3.2 Facebook zombie profiles

The next step was trying to expand our dataset of unmatched Facebook profiles. In this step, 2 zombie profiles were created who sent out a total of 200 friendship requests to the remaining 39 unmatched Facebook names.

One zombie profile sent out 114 friendship requests and got 8 accepts, a total of 7%.

The second zombie profile sent out 86 friendship requests and got 5 accepts, a total of 6%.

Of the 13 accepted friendship relations, 0 were a match with our LinkedIn subset.

4.3.3 Facebook crawling results

Our Facebook crawler was set to crawl the profiles of all the 46 employees in our LinkedIn subset of whom a definite match was found on Facebook. Since we did our initial crawl with a zombie profile on Facebook that did not have any friends, we were able to be sure to only get the publicly available data from the employees. After the iClone profile had received the final amount of 6 friend connections, these 6 profiles were successfully crawled for private data.

Of the 46 employees, 6 had their profile completely blocked off from the public.

The number of times we publicly crawled each field of a Facebook profile can be seen in table 11 in appendix B.

4.4 Summary of the final subset

The final dataset analyzed by us consisted of 86 entries. Of these 86 profiles we collected the full LinkedIn data. Out of the 86 entries, 46 were found and matched on Facebook using our matching algorithm. Of these 46 Facebook profiles, we collected the full Facebook profile information of 6 users. Of the remaining 40 profiles, our crawler was able to collect at least one field of public information (apart from name and profile picture) of 34 profiles. The last 6 profiles had their Facebook privacy settings set to not to display anything publicly. Table 9 gives an overview of the amount of data that is collected on both OSNs.

OSN	Profiles collected	Full info	Public info	No info
LinkedIn	86	86	n/a	n/a
Facebook	46	6	34	6

Table 9: Amount of data collected from the final subset on both OSNs

4.5 Theoretical research in the form of interviews and a literary study

We performed a total of three interviews with IT Security specialists, as found in the appendix section C. All three interviewees are information security specialists with years of extensive experience in social engineering. All three interviews answer a set of questions that aid us in identifying which risks companies face by our research. Furthermore, conducting these interviews helps us in locating mitigation techniques that can reduce the effectiveness of our research.

During our literary research we came across two papers we believe are the most helpful in answering parts of our subquestions. These papers, respectively [24] and[25], are used during the analysis in sections 5.2 and 5.3.

5 Analysis of the research

The results obtained by our research, as presented in section 4, are used to create an analysis to illustrate the meaning of these numbers. We present a step-by-step overview of how we got from the input data to the final output data and show how all the different components, as explained in section 3, work together to form a unified and visualized dataset.

Next, the data is interpreted in such a way that possible social engineering attacks can be constructed with the available dataset. Also, we incorporate parts of the interviews to illustrate real-life experiences with data obtained from social media for social engineering purposes.

Lastly, we present possible actions of reducing the severity and seriousness that companies can undertake to protect themselves from sort-like attacks.

5.1 Automation of hierarchically visualizing the company structure

5.1.1 Defining the subset of employees on LinkedIn

The first step of our attack scenario was a manual action in which we created a profile on LinkedIn and populated its personal and professional details with trustworthy information. The professional information was chosen carefully so that the profile would logically correspond to the standards set by the company we targeted. As previously explained in section 3.1, the 'Followers' page of the company was the first bit of information we needed to obtain. With our program, we parsed the content of the public profiles of all the followers of our company. This resulted in a total of 66 employees of whom we collected their full first name and surname. Because our fake profile initially didn't have a connection to anyone on LinkedIn, which obfuscated our search results, connection invitations were sent out to all 66 followers. Our goal was not to get a maximum amount of connections on LinkedIn, but rather find an entry point in the company's circle of employees. With the tiered-layout LinkedIn employs, the personal network of our profile would expand exponentially if connections could be made. It took less than one hour for the first connection to accept us into their personal network. The same day, 29 out of these 66 connections accepted our connection requests.

Now that our profile has expanded its professional network to the point that searching for the company name on LinkedIn wouldn't yield obfuscated results any longer, we initiated a search for employees that work at the company. With the limitations of LinkedIn's basic account, restricting the total amount of search results to 100, we got to a total of 106 invitations being sent out. This meant that of the 66 people following the company, 60 identical matches showed up in the search results (and were ignored by our program) and 40 new people were found and approached. At the end of our research our profile was connected with 39 employees of the company, out of 106 connection invitations sent.

Now, LinkedIn calculates the amount of people your profile is connected to through your 1^{st} tier connections. For our profile, this were 11,420 connections. We needed to define when we were connected to enough people to ensure that everybody in our company was approachable, at least as a 2^{nd} tier connection. We did this by reasoning that if the search results that came back after all the necessary filters were applied were at least a 2^{nd} tier connection, every employee we needed was approachable. In the case that there were still 3^{rd} tier connections or Out Of Network connections returned, more 2^{nd} tier connections would have to be connected with.

Crawling the total amount of results that were returned on our initial company search lead to a dataset that was not specific enough for our final needs. A total of 4,096 search results got returned without any filtering done. We decided not to target an entire company, but rather focus on a specific division inside this company. Doing so created a much more compact dataset to work with. To refine these search results, several filters were put in place; search only for people that live in a distance radius of 50 miles around the company and who work for the specific company division. The end result was a set of 525 employees that conformed to our criteria. The public information of these profiles was crawled.

Of these 525 employees that were found, a closer filtering was done by selecting only the 2^{nd} tier connections (with our profile) that *currently* work at the targeted company. The 2^{nd} tier connections were selected because we already crawled the 1^{st} tier connections and the 3^{rd} tier connections remained partly obfuscated. This resulted in 286 remaining profiles of which we had the full first name, the full last name and at least the company where the user currently works.

5.1.2 Using the LinkedIn dataset on Facebook

With this information, the matching algorithm (as depicted in figure 13) was employed to find possible matches between the LinkedIn dataset and all the users on Facebook. Note that to perform the profile searches on Facebook, a new profile was used that didn't have any friend relations or profile content. Figure 5 shows the combination of fields that were most frequently used to perform successful matching of a Facebook profile with a LinkedIn profile. We see that the combination of 'Current Employment' and 'Education' helped us to find a definite match 28% of the time. Including the 'Living Location' added another 20% to that list.

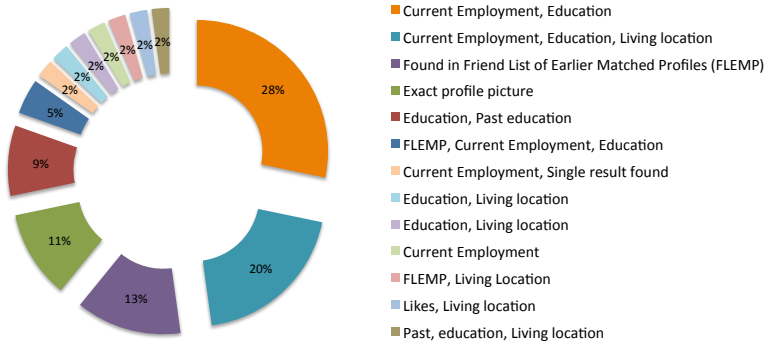


Figure 5: An overview of the combination of fields that were most frequently used to perform successful matching of 2 profiles.

Because searching for a specific name on Facebook can potentially return multiple results, the result set was split up in two lists:

- A list of employees that can be found on both Facebook and LinkedIn with a definite match
- A list of users that did not return a definite match after our matching

The list of definite matches was made up using previously researched matching techniques, plus our own criteria. Specifically we matched the first name and the last name, and then a combina-

tion of living location, profile picture, current work situation and/or education. Looking at these fields and comparing them to the information in our LinkedIn dataset, resulted in 125 LinkedIn users with a definite match on Facebook.

The other list is a list of people we could not positively identify, which doesn't necessarily mean that they don't have an account on Facebook.

For the purpose of this research, a dataset of 286 profiles was rather superfluous because proving the feasibility of aggregating techniques could be done on a smaller set of data as well. This is why we applied a last set of filters on the current work position of our entries, which redefined the dataset to 85 employees. Out of these 85 employees, 37 were in the first list and thus matched on Facebook.

In order to try and reduce the list of unmatched profiles, two techniques were used separately: FLEMP and zombie profiles, described respectively in sections 5.1.3 and 5.1.4.

5.1.3 FLEMP

Of all the 37 Facebook profiles we already identified, we searched for profiles that gave public access to their friends list. Using FLEMP we were able to identify another 9 profiles, resulting in a total of 46 Facebook profiles out of 85 LinkedIn profiles. Interestingly, this means that FLEMP helped us to find a match 13% of the time, as seen in figure 5.

5.1.4 Zombie Profiles

In an attempt to further reduce our list of unmatched profiles, we employed a technique using zombie profiles. We started by creating a first zombie profile which is populated with random data; a name generated by `naamgenerator` and likes, interests and education filled in randomly. This one profile started with searching the first name on our unmatched profiles list on Facebook, and sent friend requests to all occurrences in the results list. Then it moved on to the second name and so on until Facebook finally temporarily blocked the account. As figure 6 illustrates, Facebook detected that a high number of friend requests were sent out. At this point, we could still advance after providing a cell phone number and filling in a confirmation code. However, we decided to create a second zombie profile and continue friending the remainder of the list with that profile.

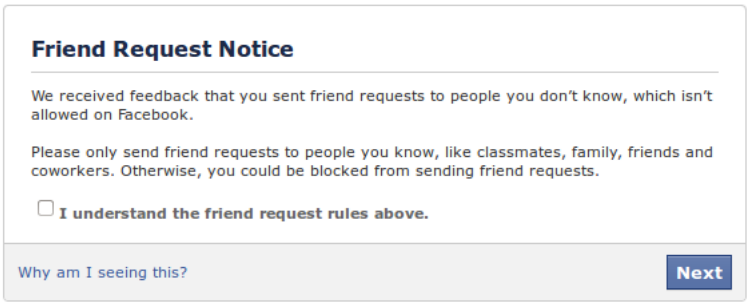


Figure 6: A notice received by Facebook after sending too many friendship requests

In total, our 2 zombie accounts sent out exactly 200 friendship requests: one to every user whose name was a direct match with our unmatched list. Of these 200 friendship requests, 13 got accepted. Unfortunately, none of the 13 profiles were a match with our LinkedIn dataset after inspecting the other fields that now became accessible.

5.1.5 Finding a suitable profile for iCloner

With our final set of 46 matched Facebook profiles, we went back to LinkedIn and initiated our iCloning technique. In order to find an account on LinkedIn that was suitable for iCloner, our matching algorithm looked at the 2^{nd} tier connections on LinkedIn and tried to find a correlation between the people through whom our fake LinkedIn profile was connected with. In order to find a user we had as a 1^{st} tier connection on LinkedIn that did not have an account on Facebook, the 'How you're connected to' feature on LinkedIn was used. The algorithm resulted in 1 account being frequently connected to 2^{nd} tier users on LinkedIn and that didn't appear in our Facebook matches database.

With this one LinkedIn account in hand, we went to Facebook and created an account filled in with all the same data as this account has filled in on LinkedIn. We copied his name, profile picture, past education, current employment and past employment. The end result was an account that looked authentic enough for it to be the real person. Next, all the connections this person had on LinkedIn and who were also on our list of matched Facebook profiles were sent a friendship request.

In total, 10 friendship requests were sent and 6 of them were accepted. Of these 6 profiles, the full private information was crawled and stored.

5.1.6 Analysis of the fake profile techniques

During our research, we used three different techniques and 4 fake profiles in order to try and get data from social networks.

The fake profile we used on LinkedIn proved to be highly effective, with a success rate of 36,7% it proves that users on LinkedIn are not too hesitant to add connections to their professional network. Making the profile appear credible to employees of a company, by filling its profile fields with company related information, will help with making the target less wary in accepting the profile.

We believe that the lack of personal data that is shared on LinkedIn is the biggest reason for users not to be too suspicious, or worried, when it comes to opening up their network. We noticed that we did get more in-depth information on the profiles that added us, but the nature of LinkedIn restricts that data to be anything other than professional information. However, the high amount of professional data that was crawled in the end made our Facebook matching easier. On several occasions were similarities among profiles on LinkedIn and Facebook found by looking at fields like education, likes, interests etc. Links that we would not have been able to make if we didn't possess the details that were put on LinkedIn.

Generally, information found on Facebook is much more personal and dynamic than on LinkedIn. The avidness with which its users use the social medium make it a source of information that is indefinably valuable. In our first approach we employed two zombie profiles that showed a relatively low success rate. We believe that is mainly due to the fact that Facebook users don't

generally accept friendship requests from strangers because it won't have any positive effect on them. In contrary, on LinkedIn, it can be assumed that expanding the professional network will make for better networking to be possible. Information on LinkedIn is often used as a digital business card whereas information on Facebook is generally only shared with friends or acquaintances.

This last statement is proven with the cloned profile. It showed that once a profile appeared authentic and targeted friendship requests were sent, the success rate is as high as 60%. Combined with that high success rate, the effect of 'Reverse Social Engineering' is not to be underestimated. We only sent 10 friendship requests yet received 4 friendship requests from other people. These were all employees of our target company and the reason why specifically employees were requesting our friendship, was because of the combination of relevant information that was provided in our cloned profile for them, and because a mutual friendship was constructed through one of the 6 established friendships.

Because of the success of the cloned profile, we deem it possible for this technique to be reproduced on a large scale, in order to fully enhance the dataset. Furthermore, a test on a larger scale has already been conducted in [18] and their final results had a similar success rate as ours, with a 56% acceptance rate for their cloned profiles.

The success rate of our aforementioned techniques is illustrated in bar chart 7, which clearly shows how the iCloner technique in percentage excels the other techniques.

5.1.7 Analysis and visualization of the aggregated data

Now that we explained how we got to all of our data, it's time to put it to show. Using the different zombie profile techniques provided us with access to profile data that would otherwise have been hidden to us. As section 2.1 explained, we will specifically be targeting service data, disclosed data, entrusted data and incidental data. The data analysis is done on the total subset of 86 profiles.

From LinkedIn, there are information fields that have been gathered with a 100% certainty. Because we've crawled the data from all our 1^{st} and 2^{nd} tier connections, all the disclosed data has been collected. Table 1 shows us which information can be collected from LinkedIn using a basic account and up to what tier we would have to go to collect as much data as possible. Apart from the connections list of the 2^{nd} tier connections, all the disclosed data was gathered.

To make a LinkedIn account a user is asked to provide service data to the OSN: his real name, e-mail address, location, job and industry. We were able to collect all of this data, apart from e-mail addresses of 2^{nd} tier connections, which are always obfuscated unless they're placed somewhere but in the therefore designated field.

The lack of privacy options for the disclosed data on LinkedIn made it very straightforward for us to crawl this information. Table 5 shows which fields we could crawl and bar chart 8 depicts how much information we managed to gather in the end and also gives an idea of how LinkedIn users, on average, fill in their profile.

Information-wise there were two things that LinkedIn lacked for our research: entrusted information and incidental information. Even though LinkedIn introduced 'Activity broadcasts' - a feature that can send status messages to your connections -, the business-oriented nature of the OSN prevents information to be shared that is considered too private. This is why we turned to Facebook, being the largest OSN and which focusses primarily on the individual, provided us

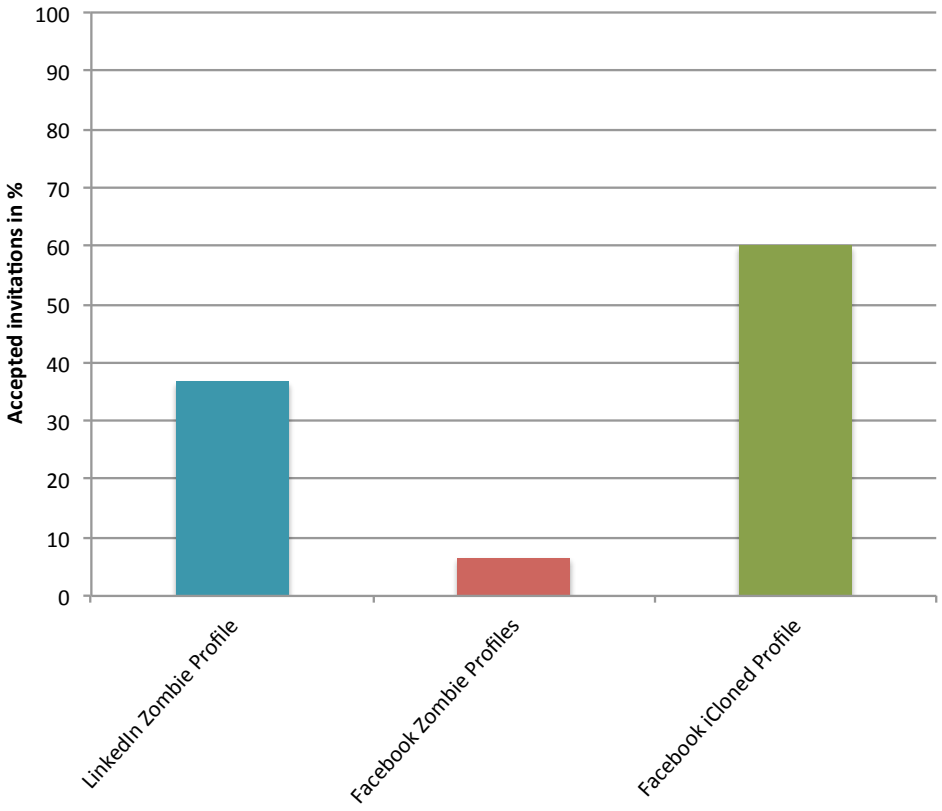


Figure 7: Accepted invitations per profile technique in percentage. The LinkedIn zombie account sent 106 invitations, the Facebook zombie profiles sent 200 invitations and the iCloner profile sent 10 invitations.

with the missing link in data. Facebook also requires its users to provide them with service data, yet it's the same as on LinkedIn, apart from the user's gender. Also, as table 3 shows, Facebook offers its users the possibility to put in a lot more disclosed information than LinkedIn does. The data enrichment aspect of our research benefited from this extended set of disclosed information and Facebook's default privacy settings didn't prove to be much of an obstruction in getting it either (table 3).

From Facebook we collected three parts of information; we started by crawling all 46 profiles publicly, just to see how much information we were able to gather. In table 11 in appendix B we show how much information we were able to gather from all our profiles. Because we managed to get access to 6 profiles using our iCloner technique, their private information was gathered too.

Finally, we looked at how much entrusted and incidental information we gathered by doing public crawling and after our friending techniques. The wall (or Timeline) is the only source where we would be able to find this information. The entrusted information would be the information the users posts on his own wall and the incidental information is the information that was posted on his wall.

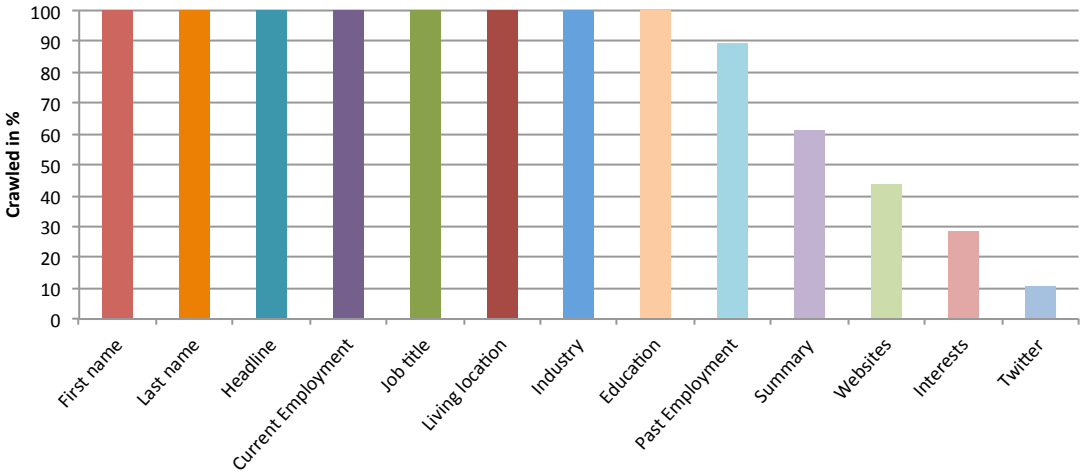


Figure 8: Overview of the amount of times LinkedIn users filled in specific information fields on their profile. The total percentage is calculated on 86 profiles.

In our data set, the wall was publicly accessible from 13 profiles out of 46. That's 28% of business professionals who left their entrusted and incidental information open for the world to see. Add up our 6 profiles that we were able to infiltrate and the percentage rises to 41%. However, we did not manage to get access to more walls posts through Facebook's 'Friends of Friends' privacy setting when looking from our iCloned profile. We believe this is because of the limited amount of people we friended in the first place and that it can still be considered viable that more walls will become accessible as the friends-base expands more. Note that we only looked at the walls of Facebook profiles that appear in our original dataset of 46 profiles and that it is possible that other users had their wall's privacy settings differently, but they weren't of interest in our research. Figure 9 gives a full overview of the amount of times we were able to crawl the Facebook fields.

After gathering all the service, disclosed, entrusted and incidental information, we were able to visualize all this data in a uniform way. In order to do this we heavily modified the open-source application Vizster to have it conform to our needs. We used a weight-ranking procedure as explained in 3.4 to build-up our graph to have the users (depicted as individual nodes) centralize around their respective position inside the company. This position attribute was gathered through LinkedIn.

Vizster allows to click on a user to have its personal information expanded and displayed alongside of the graph. At the center of our graph we find the company name and directly connected to the company are the different subdivisions we found inside the division we targeted. From each subdivision departs a hierarchical tree that shows the different positions that are held in the subdivision, as separate nodes. Each position-node has its own subset of employees that all share the same position in the subdivision.

Figure 10 depicts how the information that is gathered shows up on the right side of the graph per individual node. Note that all the values are obfuscated as MD5 sums and that all the names are generic. The job positions, however, are genuinely generated from actual LinkedIn data. It also shows how the company acts as a central node from which all the sub-divisions

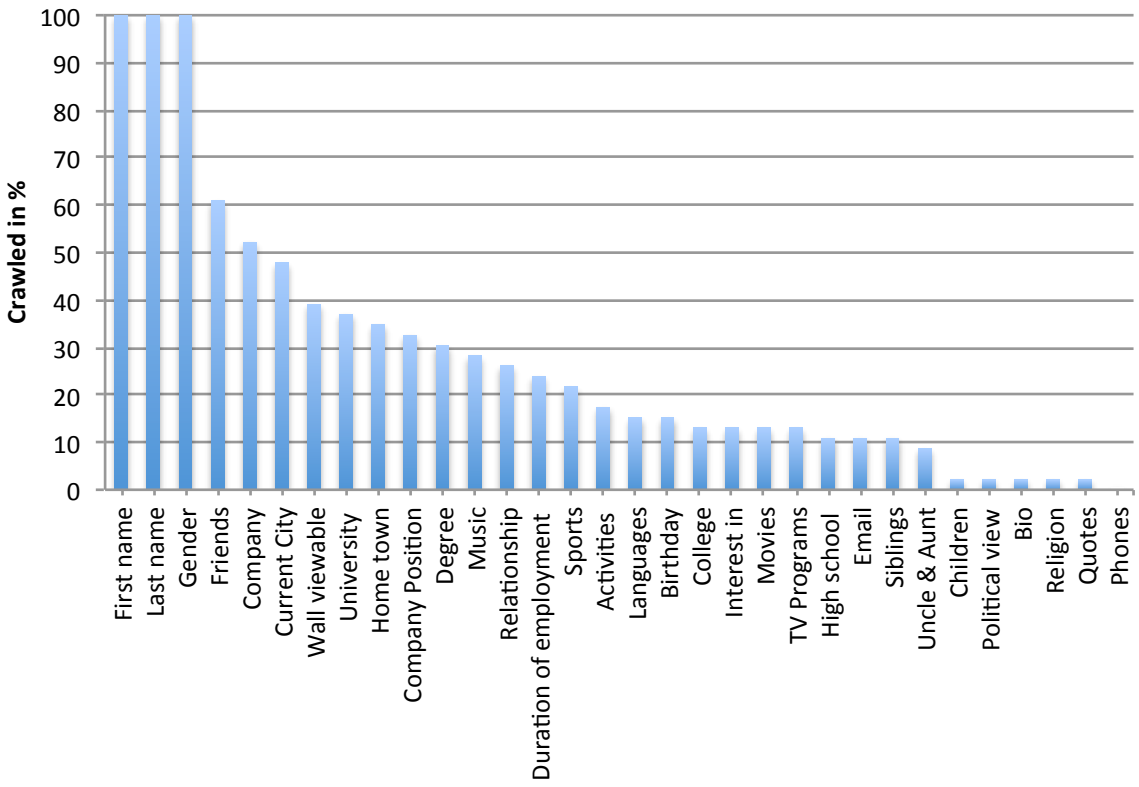


Figure 9: Overview of the amount of times Facebook users filled in specific information fields on their profile. The total percentage is calculated on 46 profiles.

spawn. Figure 11 shows our full company division, all 86 profiles divided into their job title and company sub-division.

5.1.8 Do some FOCA

One field of information that has - knowingly - not been handled yet, are e-mail addresses. We noticed that e-mail addresses were generally not displayed on both of our OSNs. Apart from the few we managed to gather through our 1^{st} tier connections on LinkedIn, we experienced a general lack of e-mail addresses from the greater part of our data set.

Because getting these addresses through the OSNs didn't appear to be feasible, we resorted to a tool called 'FOCA'[4]. FOCA is an information gathering tool that has, amongst its vast amount of features, the possibility of searching Google and Bing for any type of documents that have a relation to a certain web domain. In detail, it gathers metadata (inter alia e-mail addresses) from these documents and gives an overview of any information that can be found from these documents. It is a fair assumption that companies use the same structure for all the

[4]http://www.informatica64.com/foca.aspx

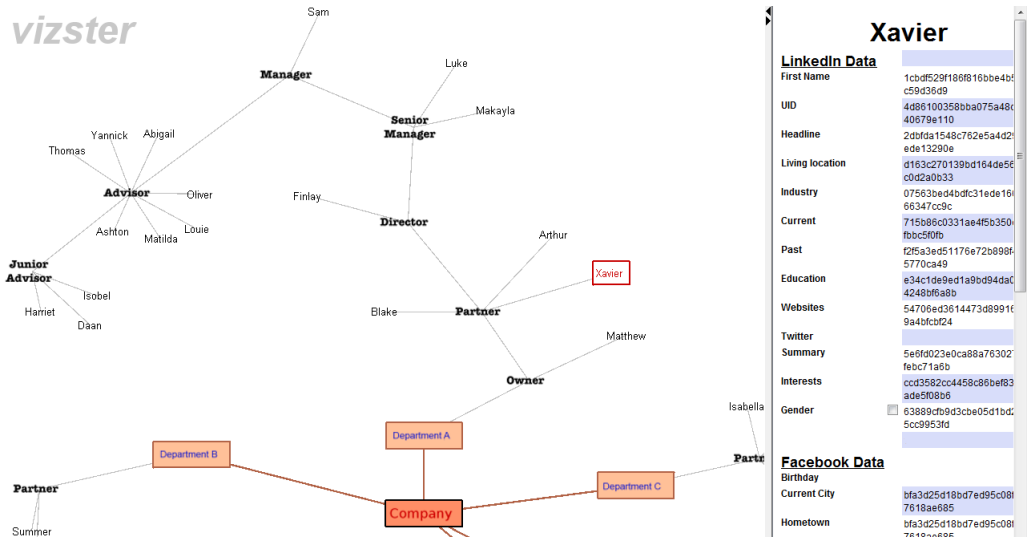


Figure 10: An example preview of how our data is visualized with MD5 obfuscated data.

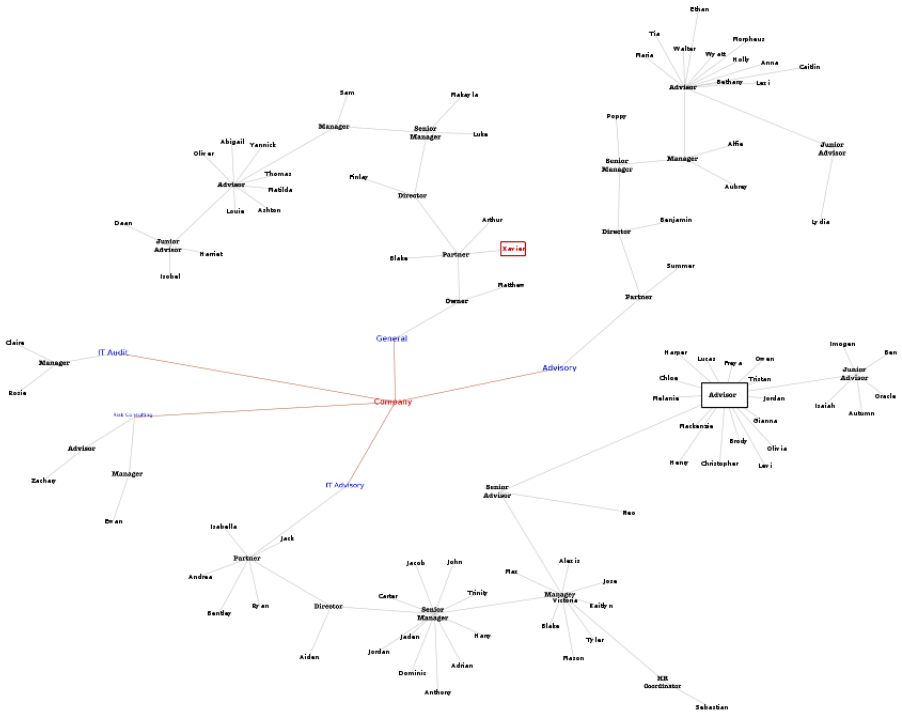


Figure 11: An example of a full overview of our 86-node large network

e-mail addresses of its employees and most of the time it involves the employee's first and last name.

By submitting a company's domain name in FOCA, it became possible to get a few e-mail addresses of employees that have ever uploaded a PDF document or an Excel worksheet for example on the web. Once the structure of the e-mail addresses was known, a full list of company e-mail addresses for all the employees in our data set could be generated.

For our target company, the structure: 'lastname.firstname@company.com' is used. For integrity reasons, screenshots from FOCA have been omitted from the report.

An added value that FOCA has provided is that once we had the company e-mail addresses we could query Facebook again and see if any profiles turned up when searching by providing the e-mail address. If the exact e-mail address is known, the Facebook search will result with exact and unique profiles that have used that e-mail address to register at the OSN. For our target company, out of 46 Facebook users, none of them used their company e-mail address to register at the OSN.

5.2 Consequences for companies

During this section we discuss the dangers that our research poses for companies. We go into detail about the amounts of data we gathered as well as the time it took. We also discuss the use of the graphical overview of the company hierarchy. We end this chapter by discussing how our research greatly benefits creating social engineering attack scenarios and finalize with showing a few scenario examples.

5.2.1 More, faster and easier data gathering

Our results show that we are able to gather a lot of information, both from LinkedIn and Facebook. In addition, the results also show that it only took us two days to collect this amount of data. These two results immediately show the advantage of our research. Namely, that gathering data from multiple sources can be aggregated and automated. The fact that we are able to automate this process makes it a lot easier to gather more detailed information, faster. A fully functioning program would allow you to only specify the name of a company along with a few keywords after which it would do all the searching, filtering, crawling and organizing of the data.

Matthieu Paques mentions in his interview, appendix C, that it takes him about three quarters of his time to prepare for a social engineering attack. In our interview with Mark Bergman (appendix C) it becomes clear that the amount of time spent on collecting information is related to the amount of potential social engineering targets. We argue that our method is able to collect a far more comprehensive dataset in a much shorter time than when done manually. We believe that our research can save social engineers huge amounts of time and work during the preparation of a social engineering attack

5.2.2 Hierarchical overview of the company structure

Jeroen Van Beek discusses in his interview, appendix C, that the use of organizational charts is in most cases not very useful. Mainly because, especially in larger layered companies, organizational

charts that can be found are often only a vague representation of the top of the company and that it lacks the detail needed to be useful in a social engineering attack. We believe that our graphical representation can serve as a great addition to, or even replace, an organizational chart. We provide a highly detailed visualization of the hierarchical relations between all the employees and their personal information, which can be just the level of detail needed to be useful in a social engineering attack.

Furthermore, very specific types of social engineering attacks can be greatly improved by our visualization. These attack, explained in more detail in [25], are based on creating a false sense of authority by referencing people placed higher in the company hierarchy or by making a personal connection with the victim. We believe that the level of detail presented in our overview can potentially lead to a more precise and effective execution of these types of attacks.

5.2.3 Creating social engineering attack scenarios more easily

Another interesting aspect of our research is the fact that you have a central place where all the data is aggregated and stored. This means that any number and combination of filters and search queries can be performed on the dataset. This allows a social engineer to search for very specific types of information, information that is needed to prepare social engineering attack scenarios as mentioned by Matthieu Paques and Mark Bergman in their interviews. We believe that by making this data so easily accessible it becomes easier for social engineers to define attack scenarios.

To strengthen this belief, three social engineering attack scenarios have been selected that show how our research could be used to aid in the process of preparing an attack scenario. All three scenarios have actually taken place in practice and are discussed either by Matthieu Paques in [25] or during the interviews shown in appendix C.

Scenario 1: Making a personal connection

Psychological 'tricks' are regularly used by social engineers to get to information that they normally wouldn't get access to[25]. A very effective trick is trying to make a personal connection with the victim. This is often done by engaging in small talk and being able to talk about similar interests, problems or experiences. A social engineer can literally build a bond of trust between the victim and himself by indicating that, for example, they went to the same school, had the same printer problem a couple of times, went to the same hotel on a holiday, bought the same book and so on. It will become harder for the victim to refuse a request for information after this bond of trust has been established. The detailed and personal data that our research provides is the perfect source to get to the information needed to make this personal connection.

Scenario 2: Targeting a specific group of employees

In a scenario described by Matthieu Paques in [25] a fake online 'give away' contest is set up. He targets young female employees as to not accidentally target members of the IT department or higher placed managers (people who should know if such a contest was given). By means of a specially crafted website using the company style, the employees are asked to fill in their credentials before they could enter the contest. Within half an hour he had already collected several credentials that could be used in subsequent technical attacks. The aggregation of data by our research allows a social engineer to perform exactly this kind of filtering. Since all the

data is sorted, a social engineer could easily create and execute filters along the lines of 'give me all the female employees between the ages 20 and 30'. This allows him to target very specific groups of employees and to tailor his attack scenarios accordingly.

Scenario 3: Using the company structure

Another scenario described by Matthieu Paques in [25] has the goal to obtain system access. He talks about how he he called an employee pretending to be of IT support with the message that there are network issues and that her system could crash because of it. He asked whether she also noticed that her laptop started working slower, which she confirmed it did. He then said that he would start working on the problem and hung up the phone. Two days later he called that same employee with the message that the problem had not been solved yet and that it was most likely a local problem. He asked when she would be able to bring her laptop to IT support on which she replied that she was rather busy. He then proceeded to say that he could make an exception and 'troubleshoot' the problem remotely if she would be willing to change her password to something temporary. This she did, after which he had successfully achieved the goal of the assignment. In this specific scenario it becomes clear how important it is to know who performs which function inside the organization. Having a hierarchical overview of the company greatly benefits this type of scenario and could even be the reason for choosing such a scenario. The detailed overview allows a social engineer to carefully select his targets. He also knows whether that person has any relations with, for example, the IT support department.

5.3 Mitigation of the data aggregation process

The previous section zoomed in on potential dangers for companies by showing the benefits of our research for potential social engineering attacks. This section discusses mitigation solutions that can be employed to combat the effectiveness of our research. We start by looking at what can be done to mitigate the social synergy between networks and how to reduce the effect of our data gathering techniques. We continue by explaining that being more generic on LinkedIn reduces the accuracy of our company hierarchy visualization and finally we discuss the importance of user awareness and how it can be created using our research.

5.3.1 Preventing social synergy

The effectiveness of our research and the amount of data gathered by our research can be greatly reduced when no matches between online social networks can be made. We believe that company policies that describe certain restrictions on what or what not to post on certain online social networks can be a very effective mitigation. As already explained, the fields 'Current employment' and 'Education' are most often used in matching profiles and restricting these fields to just LinkedIn would already significantly decrease the amount of profiles that can be matched. An alternative but less effective approach could be to configure the privacy settings in such a way that the fields 'Current employment' and 'Education' are not publicly visible on Facebook. However, this still means a match could potentially be made based on these fields when zombie profiles or iCloner techniques are successfully used.

Based on the graph shown in figure 5 we can at least define the recommendations as shown in table 10. This table shows for each applicable field used during the matching process a very basic

recommendation on how to mitigate its use for social synergy. Of course, when additional fields are used for matching, additional recommendations would need to be defined.

Field	Recommendation
Current Employment	Restrict to LinkedIn
Education	Restrict to LinkedIn
Past education	Restrict to LinkedIn
Likes/Interests	Restrict to Facebook
Profile picture	Use different profile picture on each OSN

Table 10: Recommendations to reduce the effectiveness of social synergy

The effectiveness of matching profiles using the FLEMP technique is directly affected by the aforementioned recommendations. The effectiveness of this technique is solely dependent on the public availability of the friends lists of earlier matched profiles. Thus, by reducing the effectiveness of matching profiles with public data the effectiveness of FLEMP is also automatically reduced. If you would also disable the public availability of your friend list, we believe this technique will become harder to use successfully.

5.3.2 Reduce the effect of data gathering techniques

The techniques used in our research both aim at collecting public and private data. The only way to prevent the collection of public data is by not having it publicly available. This directly relates to being aware of which information is public and how to control this by using the privacy settings of the specific OSN. How to control these privacy settings has been thoroughly discussed in sections 2.2 and 2.3.

We have shown that we make use of zombie profiles and iCloner profiles to try and collect as much private information as possible. Of course, this is only possible when invitations sent from these two types of profiles are accepted. In order to prevent the collection of private data it is important for users to be sure that they only accept invitation requests from people they know. Based on the insights we gained during our research we have defined three points of interests which, if followed, reduce the effect of our data gathering techniques.

1. On LinkedIn, only add people you have spoken to or had business with in the past few days.
2. Be suspicious of invitations coming from people who have often claimed or mentioned that they do not use that specific OSN.
3. Never accept invitations requests of people you do not know.

5.3.3 Be generic on LinkedIn

The visualization of a company's hierarchy as displayed in our results is constructed solely from data collected on LinkedIn. As explained in section 3.4, we built this visualization based on the 'Current employment' field as filled in by the users on their LinkedIn profile. We believe that a possible solution for mitigation could be to be more generic on LinkedIn. By omitting certain details from the 'Current employment' field, the accuracy of the visualization will decrease. This is better explained with an example in which a person has the 'Current employment' field filled in

with his job title 'Senior database engineer at IT-Solutions4all engineering'. Our technique can already distil three important pieces of information. First, the person is a senior engineer. Second, the person works at 'IT-Solutions4all' and third, he works at the engineering department. If this same person would fill out the 'Current employment' field with 'Engineer at IT-Solutions4all', we would be unable to distil as much information as before. The more people who do this, the less accurate our hierarchical overview becomes and how less useful it becomes for a social engineer.

Of course, specifying exactly what it is that you do and for whom you do it is an integral part of LinkedIn. Therefore, being generic might not be seen as a real solution for some people. This then becomes more of an issue on how important it is for a company to not have its hierarchy visualized. In turn, these are policies and decisions that need to be made by each specific company themselves.

5.3.4 Employee training and creating user awareness

Defining and implementing policies can only work if employees follow them. It is important for users to be aware of what can be done when their data is aggregated, organized and visualized. In order to make the risks clear to employees, the first thing that needs to happen is that they need to be educated to make them less prone to social engineering attacks. Employees can be trained to be wary in certain uncommon situations, they can be taught to detect typical social engineering attacks by giving real-life examples or role-playing games. Secondly, Jeroen van Beek mentions in his interview that giving compulsory trainings hardly works and that people need to be put in an 'open' mindset before they are willing and able to change. This open mindset can be achieved by holding awareness sessions in which you are able to find one person who is convinced that nothing can ever happen to him, and then show him that you breached his privacy. The effect of such awareness training sessions become very effective when performed in a fun, realistic and interactive way.

With this in mind, we believe that it is valuable for companies to periodically test what information can be gathered online with our research methods. Information that pertains not only to the company itself but also to its employees. If this is followed by awareness sessions with very concrete examples, we believe employees will become more aware of the dangers and are most likely more willing to conform to company defined policies regarding the use of OSNs.

6 Conclusion

During this section we state our conclusions as well as the reasoning behind them. We divided this section into multiple subsections, each answering one of our research questions. Lastly, a final conclusion is derived from the experiences and knowledge gathered during this project.

All our research questions can be seen in section 1.2. However, for the purpose of clarity we mention each research question in their respective section.

6.1 Successfully combining information gathering techniques

Our first research question, *'How can current information gathering techniques be combined to achieve this goal?'*, can be answered by looking at the defined flowcharts in figures 12, 13 and 14 in appendix A. These flowcharts show a step-by-step decision-making process which starts by crawling user data on LinkedIn, continues by matching LinkedIn users to their Facebook profiles and ends with crawling users' data on Facebook.

Our flowcharts show when to use which type of OSN profile in each stage of the information gathering process, how to identify LinkedIn profiles on Facebook and finally how and when to crawl information on Facebook and LinkedIn. By automating these steps we have shown that current information gathering techniques can be combined to achieve our goal. We not only conclude that all three information gathering techniques can be combined, but also that they can be used to great effect.

6.2 The consequences for companies

The second research question, *'What are the consequences for companies?'*, can be answered by looking at the results in section 4 and the analysis in section 5.2. Based on this analysis we conclude that the main risk for companies is that social engineers can gather a vast amount of employee information a lot faster using our research. In only two days of sending invitations and crawling data, we have shown the wealth of information that can already be collected.

It also becomes easier for a social engineer to create attack scenarios. We have already given a few example scenarios in section 5.2.3 that show the dangers of the data collected during our research. In addition, the information is now stored in a way that allows for far more comprehensive search queries to take place.

Finally, we conclude that the use of a visualized company hierarchy can greatly benefit a social engineer. As explained in section 5.2.2, our visualization is more detailed when compared to organizational charts provided by companies. This makes our visualization particularly effective when attacks make use of 'referencing people placed higher in the company's hierarchy', in order to create a false sense of authority.

6.3 Mitigation procedures for companies

Multiple conclusions for our third research question, *'What can companies do to mitigate this process?'*, can be defined. First off, we conclude that the main technique to make matching LinkedIn profiles on Facebook less effective, is to prevent social synergy. By defining the way

information should be shared on both OSNs, as stated in section 5.3, we believe that the social synergy between the OSNs will be greatly reduced, resulting in less information that can be gathered.

Mainly by configuring Facebook's privacy settings in such a way that less information becomes publicly available, it becomes harder to match LinkedIn profiles to their Facebook counterparts using only public data. In addition, shielding the friend list from the public reduces the effectiveness of our FLEMP technique, making it even harder to find matches.

We also propose three guidelines (section 5.3.2) for users to decide when or when not to accept connection or friend requests. If these three points are followed, we conclude that crawling private data becomes almost impossible.

Next to defining policies and guidelines for using social media, companies can educate their employees on the dangers of social engineering and would benefit greatly if they would use our research to periodically test the available information that can be found on their employees. With the information they gather they can hold awareness sessions for their employees to stimulate general user awareness.

6.4 Visualizing a company's hierarchy and its employees

Before continuing and answering our main research question, we specify it once more:

> *How can Online Social Networks be used in the automated creation of a graphical view of the company hierarchy and its employees for the purpose of social engineering?*

With the technologies and approaches described in section 3 we crawled, identified and matched different profiles on different OSNs. By doing so we gathered a large set of detailed information and were able to use that data to construct a graphical representation of the company hierarchy. We therefore conclude that the obtained data from different OSNs is detailed enough to create a hierarchical view of a company and to be applied for social engineering attacks.

Our results become primarily possible because of the wealth of relevant information that can be found and aggregated on multiple OSNs. All this information can be found because people are generally not aware of how much information they share online and how easy it is for people like us to access it.

7 Future work

We have proven how successful a combination of multiple OSNs can be for aggregating a large set of detailed information on a company and its employees. We believe that adding geolocation services such as Foursquare, Facebook Places or Google Latitude to the information gathering process could provide another valuable source of information for potential social engineering attacks. Because of time constraints we were not able to implement these geolocation services and therefore believe it would be interesting to see how well these services could be fitted in the information gathering process. In addition, further research could be done into adding different OSNs (Twitter, Hyves, MySpace, Google+) for an even more detailed dataset of information.

Furthermore, a business-oriented OSN such as Yammer could be used to gather more inside-information on the targeted company. Not only would detailed information for each employee be available, also detailed company information would be available. We therefore state that it would be interesting to research the added value of using business-oriented OSNs for the data-enrichment process.

In section 2.1 we explained the 4 types of data that we would be gathering during our research. Schneier also defines 'behavioral data' and 'derived data' as types of data that can be collected on OSNs, data that we have not been able to collect due to the timespan and multiple recordings needed to correctly analyze this information. The new Timeline feature that Facebook offers provides a view of a person's activity in a chronological order. Right now, our research collects all the data disregarding the chronological order in which it happened. Additional research could be done into whether the chronological order of a person's Facebook activities could be preserved and if the newly collected data types can be successfully used in social engineering attacks.

8 Acknowledgements

We would like to greatly thank Marc Smeets and Marek Kuczyński for their continuous support and advice during our research. Their invaluable experience has also greatly aided us with the construction of the final report.

We would also like to thank Matthieu Paques, Mark Bergman and Jeroen Van Beek for allowing us to interview them and for sharing their experience and knowledge.

Lastly, we would like to thank Cees de Laat for making the research projects in their current form possible.

References

[1] K. D. Mitnick and W. L. Simon, *The Art of Deception: Controlling the Human Element of Security*. New York, NY, USA: John Wiley & Sons, Inc., 2003.

[2] I. S. Winkler and B. Dealy, "Information security technology?...don't rely on it: a case study in social engineering," in *Proceedings of the 5th conference on USENIX UNIX Security Symposium - Volume 5*, SSYM'95, (Berkeley, CA, USA), pp. 1–1, USENIX Association, 1995.

[3] A. Dolan, "Social engineering," SANS Institute InfoSec Reading Room, SANS Institute InfoSec Reading Room, 2004.

[4] S. van Belleghem, "Social media around the world," tech. rep., 2010. http://www.slideshare.net/stevenvanbelleghem/social-networks-around-the-world-2010.

[5] C. Hadnagy, *Social Engineering: The Art of Human Hacking*. John Wiley & Sons, November 2010.

[6] B. Schneier, August 2010. http://www.schneier.com/essay-322.html.

[7] L. Corporation, "About us," 2012. http://press.linkedin.com/about.

[8] Alexa, "Linkedin's statistics information," 2012. http://www.alexa.com/siteinfo/linkedin.com.

[9] D. de Solla Price, "Networks of scientific papers," in *Science*, Vol. 149 no. 3683, pp. 510–515, 1965.

[10] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," 286 (5439), pp. 509–512, October 1999.

[11] .-L. Barabási, "Linked: how everything is connected to everything else and what it means for business, science, and everyday life," 2003.

[12] LinkedIn, "Now companies too have profiles on linkedin," March 2008. http://blog.linkedin.com/2008/03/20/company-profile/.

[13] Facebook Inc., "Facebook statistics," April 2012. http://newsroom.fb.com/content/default.aspx?NewsAreaId=22.

[14] Alexa, "Facebook ranking," Jube 2012. http://www.alexa.com/siteinfo/facebook.com.

[15] M. McKeon, "The evolution of privacy on facebook," 2010. http://mattmckeon.com/facebook-privacy/.

[16] D. P. W. PARTY, "European data protection group faults facebook for privacy setting change," 2009. http://ec.europa.eu/justice/policies/privacy/news/docs/pr_12_05_10_en.pdf.

[17] E. Petridou and M. Kuczynski, "Synergy of social networks defeats online privacy," 2011.

[18] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: automated identity theft attacks on social networks," in *Proceedings of the 18th international conference on World wide web*, WWW '09, (New York, NY, USA), pp. 551–560, ACM, 2009.

[19] J. Bonneau, J. Anderson, and G. Danezis, "Prying data out of a social network," in *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, ASONAM '09, (Washington, DC, USA), pp. 249–254, IEEE Computer Society, 2009.

[20] D. Irani, M. Balduzzi, D. Balzarotti, E. Kirda, and C. Pu, "Reverse social engineering attacks in online social networks," in *Proceedings of the 8th international conference on Detection of intrusions and malware, and vulnerability assessment*, DIMVA'11, (Berlin, Heidelberg), pp. 55–74, Springer-Verlag, 2011.

[21] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money," in *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, (New York, NY, USA), pp. 93–102, ACM, 2011.

[22] K. voor Mobiliteitsbeleid, "Mobiliteitsbalans 2010," 2010.

[23] M. Fowler, "Organization structures," 1996.

[24] M. Bergman, "Social engineering: An approach for risk analysis and improvements," January 2008.

[25] M. Paques, "Social engineering: The art of deception," in *Compact*, Information Security, pp. 41–48, Uitgeverij Kleine Uil, 2011.

A Flowcharts

Figure 12: Flowchart of the LinkedIn crawling process

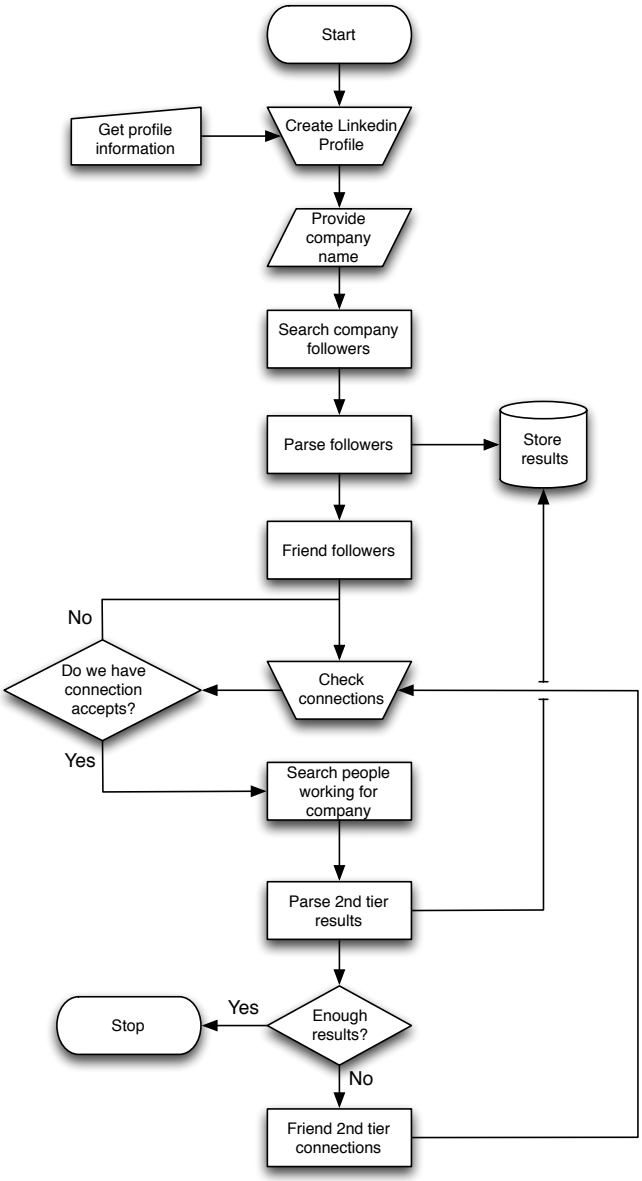


Figure 13: Flowchart of matching LinkedIn results with Facebook profiles and decision-making choices on creating what kind of profile on Facebook.

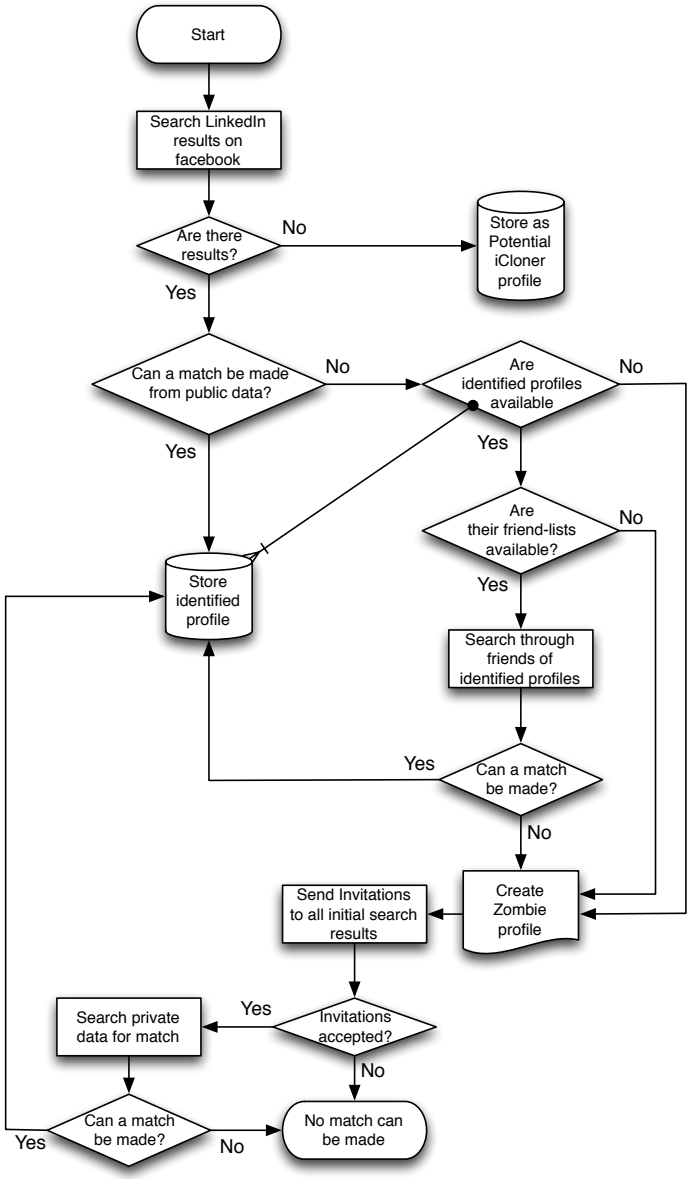
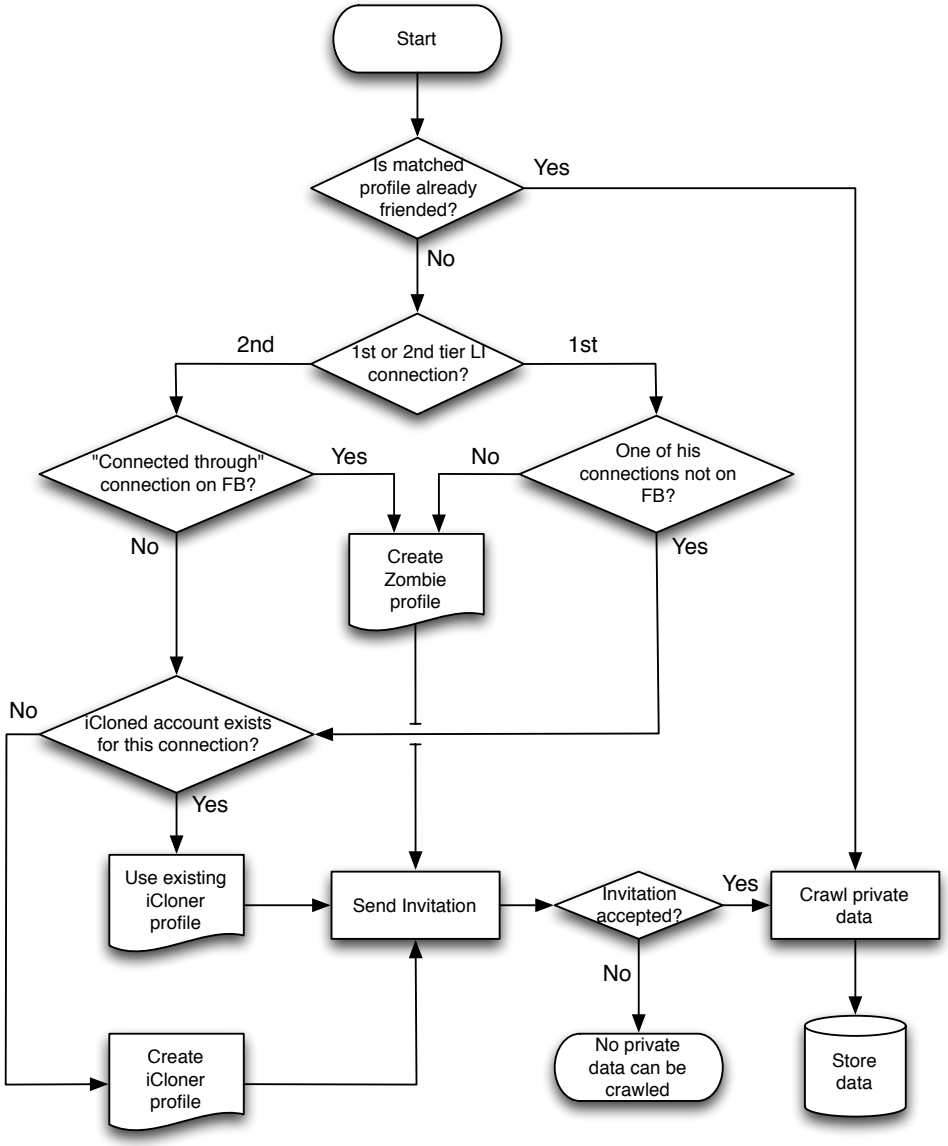


Figure 14: The private Facebook crawling process



B Tables

Field	Times crawled
First name	46
Last name	46
User id	46
University	17
Relationship	12
Children	1
College	6
Interest in	6
High school	5
Profile link	46
Languages	7
Degree	14
Email	5
Activities	8
Company	24
Phones	0
Company Position	15
Gender	46
Music	13
Duration of employment	11
Birthday	7
Movies	6
Current City	22
Sports	10
Home town	16
TV Programs	6
Political view	1
Uncle & Aunt	4
Bio	1
Religion	1
Siblings	5
Quotes	1
Wall viewable	18
Friends	28

Table 11: Number of Facebook fields crawled

Field	Times crawled
First name	85
Last name	85
User id	85
Education	85
Headline	85
Current Employment	85
Living Location	85
Industry	85
Past Employment	76
Websites	37
Twitter	9
Summary	52
Interests	24

Table 12: Number of LinkedIn fields crawled

C Interviews transcriptions

M.P.: Matthieu Paques
M.B.: Mark Bergman
J.V.B.: Jeroen Van Beek

Date of interview with Matthieu Paques: 15th of June, 2012.
Date of interview with Mark Bergman and Jeroen Van Beek: 18th of June, 2012:

Q: Which information is important for a social engineer?

M.P.: Information is always the first step. Having the first name and the surname of employees gives the possibility of forming company e-mail addresses. Companies usually employ the same construction for e-mail addresses, for example a combination of the first name and the surname. Having a large list of e-mail addresses will allow social engineers to try phishing attacks amongst others. LinkedIn is the perfect medium to find an overview of employees that work for a specific company.

It also depends on the goal of the assignment. If gaining company access is the goal, using services like Google Streetview can come in handy to scout the building, see where the cameras are and what entrances are available.

An organizational chart can also be used to find out the relations between employees, which can be relevant information.

M.B.: Names and telephone numbers and who's the boss of who. In detail, it's interesting to know the name and e-mail of the employees and what their position and function is in the company. An organizational chart would be a nice add-on but I've never used it before.

J.V.B: Knowing who knows who in a company and the relations between those employees. Knowing what each employee does in the company is also always nice to know. Having an organizational chart depends on the company. If the company is too layered, a diagram will mostly be a vague representation of the entire structure of the company and you can often be informed wrongly. Still, it's always interesting to have.

Q: Where do you obtain your information?

M.P.: Social media such as LinkedIn and Facebook and the corporate website.

M.B.: Google and just calling them. Also social media such as LinkedIn and Facebook is an easy way to obtain a bulk of information. It always helps to drive by the building and do reconnaissance.

J.V.B.: Google and social media. Also by inspecting the outside of the building in real life or in google street view. People are generally helpful but can become suspicious, so it helps to remove any doubt by talking about something mutual, such as a flagpost that is missing or broken, or education.

Q: How much time does it take you to collect this information?

M.P.: Especially for assignments where I know hardly anything about the company does it take me quite some time. Generally speaking, I reserve about three quarters of the total project time for data gathering purposes.

M.B.: It depends on the size of the company as well as how much information I have to collect. If I have to collect information about just a few persons in order for my social engineering attack to be successful, it just takes me an evening. However, if the amounts of subjects increase, it will take me a considerable more amount of time.

J.V.B.: Usually just an evening. However, I like to go in 'low tech' meaning that I pretend to be, for instance, a 'plant counter' and that I am here to count the plants in the building. Doing so allows me to know hardly anything about the people in the building.

Q: How do you typically use the information you have obtained?

M.P.: You try to create a context between yourself and the victim, for example sharing the same education gives a great topic to talk about. You can bond with the target and create a trust that will more easily remove any doubt with the victim.

You can also try and talk to an employee and act as if you had communication with his superior. This way, people can be manipulated to give you information that they would usually not give. Here is where the relations between employees become important.

Furthermore, I try to lay out every possible scenario and be prepared for anything that can happen.

M.B.: I create some scenarios beforehand but not in too much detail. In the end, it's a creative process and situations will always change, so it's important to be able to adapt to any situation.

J.V.B.: You have to see how it goes, it's always a bit embarrassing to fool people all the time, but that's something you have to get over as a social engineer.

Q: Are there people or departments that you try and avoid when trying to gain access to a building?

M.P.: I prefer to avoid Human Resources because they're usually up to date on everybody that is supposed to be working at the building at any time. They'll be suspicious when you come with a bogus story, if they have no knowledge of it.

I prefer to get access to the IT department because these users usually have elevated rights on the system. If an unlocked computer can be found, a lot of damage can be done to the systems.

M.B.: I try to avoid the less social people, but it's kind of hard to identify these. People who work in IT consulting for example are usually aware of the whole social engineering concept, so it'll be harder to fool them.

J.V.B.: I try to avoid critical people, but they can be everywhere. I don't think that any specific department is a liability, it just depends on the individuals that work there at that specific time.

Q: Are there typical weaknesses you encounter at companies?

M.P.: Wearing badges... In companies usually nobody wears their badge. If a visitor gets a visitorsbadge to get in, all he has to do is take off the badge and suddenly he's an employee. The funny thing is that no matter how much you patch and secure the outside of the building, people will always be the weakest link. The overall awareness of employees is too low.

J.V.B.: People are helpful.

M.B.: Access doors before the reception. You already have access to the building before you have to apply at the reception. Employees not wearing badges, as soon as the badge gets taken off you're a full employee. The smoke entrance is used frequently to get in the building, I don't smoke but in this case I'll happily pretend to!

Also, smoke areas. A lot of these entries are left open because people are afraid that they have to clock out for every smoke session. So they block the door to get unobstructed access to the smoke area.

J.V.B.: As soon as you're in, most access restrictions disappear.

Using the smoke entrance, just tagging along with smokers always works wonders.

I also notice that procedures are often bad. You can say, for example, 'My badge isn't working!'. Usually you'll have to give an ID but if you say that you forgot your wallet for example and only have a piece of paper with a name on it, it's usually sufficient to still get a temporary pass.

Q: What are other dangers of social engineering except trying to gain access to a building?

M.P.: Calling and e-mailing. Using phishing emails to get password credentials. We sometimes send faxes or post letters to try and create a more authenticated scam than e-mails usually are.

M.B.: For example, you can set up a new webmail and mail the employees that the first 500 registrants can win a prize. All they have to do is login and test the new system, and thus, entering their credentials.

J.V.B.: Enticing people works best by convincing them that they can get something for free. Tell them they can win an iPhone if they test a certain system and surely, a lot of people will login and give you their credentials.

Q: What kind of policies or guidelines would you advise companies to use concerning the use of social media?

M.P.: That's a difficult one... You have to keep in mind that there is a separation between work and the private life, so I think it's a bit farfetched to tell people what they can or can not publish. Of course, you can ask not to share any company related information online, but it's nearly impossible to make that concrete.

J.V.B.: Do awareness sessions. Try to find the one person who is so convinced that nothing can ever happen to him, and then show him that you breached his privacy. Those kind of awareness sessions are very effective in a company.

Avoid doing compulsory trainings, because people won't care. They'll care once they see their own face on a big screen with all their private data beneath it.

M.B.: Different people work at different departments, what I mean is that it's really hard to make overall guidelines. You could make policies per department and tell a System Administrator not to have any contact with the clients, for example.

But people have to want to change. That's why these awareness sessions are always a good idea, because it will make people realize that something is wrong.