

Submitted by
Dusan Jaguzovic

Submitted at
**Institute of Networks and
Security**

Supervisor
**Assoc. Prof. Mag. Dipl.-Ing.
Dr. Michael Sonntag**

Month Year
June 2017

Who is looking for me online?



Master Thesis

to obtain the academic degree of

Master of Science

in the Master's Program

Networks and Security

STATUTORY DECLARATION

I hereby declare that the thesis submitted is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.

This printed thesis is identical with the electronic version submitted.

Place, Date

Signature

Abstract

Is it possible to find out who is interested in you or in a particular person online? Is it possible to detect who is looking for someone's name, for someone's address or for some similar private information about a particular person, maybe on Google or Facebook? There are many reasons why someone is looking for personal information about someone else.

To identify who is interested in a particular person could have some advantages for this person. Finding out that somebody is interested in your CV could be useful when looking for a job. On the other hand, finding out that somebody is looking for your address could be very helpful if this someone is a debt collector.

Nowadays it is important to control private data which is stored or redistributed on the World Wide Web. This thesis is about finding people or machines that work with private data of an individual online. An additional goal is to have more control over this data and at the same time to have more control over privacy that is becoming more and more violated.

Kurzfassung

Ist es möglich herauszufinden wer sich für Sie bzw. für eine bestimmte Person im Internet interessiert? Ist es möglich herauszufinden wer nach dem Namen, nach der Adresse oder nach einer anderen Art privater Information über eine Person sucht, beispielsweise auf Google oder Facebook? Es existieren viele Gründe warum jemand nach persönlichen Information eines anderen sucht.

Information zu erlangen dass sich jemand für eine bestimmte Person im Internet interessiert, bringt einige Vorteile für die Person. Zu wissen dass jemand Interesse für den Lebenslauf einer Person hat kann bei der Arbeitssuche hilfreich sein. Zu wissen dass jemand nach der Wohnadresse sucht kann sehr vorteilhaft sein falls derjenige der sucht ein Schuldeneintreiber ist.

Heutzutage ist es sehr wichtig die Kontrolle über private Daten zu behalten, die in World Wide Web gespeichert und verteilt werden. Diese Arbeit beschäftigt sich mit der Suche nach Personen oder Maschinen die mit privaten Daten einer Person im Internet arbeiten. Weiteres Ziel dieser Arbeit ist es mehr Kontrolle über private Daten zu haben und gleichzeitig die Privatsphäre, die immer öfter verletzt wird, besser zu kontrollieren.

Table of contents

1	Introduction	1
1.1	Problem	2
1.2	Objective	3
1.3	Construction	4
2	Theoretical foundations	6
2.1	Current ways to find out who is looking for you	9
2.2	Information needed for client's identification	11
2.3	Reverse search	12
3	Practical Part	14
3.1	Honeypots	14
3.1.1	Honeypot specification	17
3.1.1.1	Hardware	17
3.1.1.2	Software	18
3.1.1.3	Router	18
3.1.2	Creating the honeypot website	20
3.1.2.1	Honeypot features	22
3.1.3	Honeypot availability	35
3.1.4	Tempt visitors to visit	38
3.2	Social networks	39
3.2.1	Facebook	41
3.3	Career-oriented social networks	45
3.3.1	XING	45
3.3.1.1	Identification based on user ID	47
3.3.1.2	Identification with google image search	48
3.3.2	LinkedIn	49
3.4	File Tracking	50
3.4.1	PDF Tracking	52
3.4.2	DOC(X) Tracking	55
3.5	E-Mail	57
3.5.1	Received E-Mail	57

3.5.2	Sent E-Mail	61
3.6	Tumblr	63
4	Tests	65
4.1	Honeypot Test	65
4.1.1	Interpretation of results	66
4.2	Facebook Test	68
4.3	XING Test	73
4.4	LinkedIn Test	83
4.5	E-Mail Test	84
4.5.1	Received E-Mail	84
4.5.2	Sent E-Mail	89
4.6	Test summary	90
5	Summary	91
5.1	Conclusion	93
5.2	Outlook to future	94
6	Bibliography	95

List of abbreviations

API	Application Programing Language
CV	Curriculum Vitae
DMZ	Demilitarized Zone
DNS	Domain Name System
EFF	Electronic Frontier Foundation
GPS	Global Positioning System
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IPv4	Internet Protocol Version 4
IPv6	Internet Protocol Version 6
ISP	Internet Service Provider
LAN	Local Area Network
MAC	Media Access Control
PC	Personal Computer
PDF	Portable Document Format
PHP	PHP: Hypertext Preprocessor
SMTP	Simple Mail Transfer Protocol
SMTPS	Simple Mail Transfer Protocol Secure
SSID	Service Set Identifier
TLD	Top Level Domain
URL	Uniform Resource Locator
VLAN	Virtual Local Area Network
W3C	World Wide Web Consortium
WWW	World Wide Web

List of figures

Figure 1: Google Searches.....	6
Figure 2: Personal information available online	8
Figure 3: Image search example.....	13
Figure 4: Honeypot location.....	16
Figure 5: Appearance	21
Figure 6: Canvas fingerprint	31
Figure 7: IPv6 google access.....	36
Figure 8: IPv6 deployment in Europe	36
Figure 9: Honeypot google ranking.....	38
Figure 10: Social network users	39
Figure 11: Social network login	40
Figure 12: Xing visitors	46
Figure 13: Xing example result.....	47
Figure 14: File tracking	51
Figure 15: Most popular document formats on the web	56
Figure 16: E-Mail header result.....	60
Figure 17: StatCounter platforms.....	63
Figure 18: Tumblr example result	64
Figure 19: Facebook Dusan's friend suggestions	70
Figure 20: Facebook Patrick's friend suggestion.....	71
Figure 21: Facebook Tobias's friend suggestions.....	72
Figure 22: Facebook Patrick's friend suggestions	72
Figure 23: Xing visotor entry #1	74
Figure 24: Xing visitor result #1.....	74
Figure 25: Xing visitor entry #2	74
Figure 26: Xing visitor reuslt #2.....	74
Figure 27: Xing no profile picture visitor.....	75
Figure 28: Xing visitor entry #1, method 2.....	77
Figure 29: Xing visitor result #1, method 2	77
Figure 30: Xing visitor entry #2, method 2.....	78
Figure 31: Xing visitor result 2, method 2	78

Figure 32: Xing visitor entry #3, method 2.....	79
Figure 33: Xing visitor result #3, method 2	79
Figure 34: Xing visitor entry #4, method 2.....	80
Figure 35: Xing visitor result #4, method 2	80
Figure 36: Xing external visitor.....	82
Figure 37: Who's viewed your LinkedIn profile.....	83
Figure 38: Mail Test 1 Result	84
Figure 39: XXXLutz whois result	85
Figure 40: Mail Test 2 Result 1	86
Figure 41: Mail Test 2 Result 2	87
Figure 42: Mail Test 2 Result 3	88
Figure 43: Used E-Mail app	88
Figure 44: Public IP Address belongs to certain provider	88

List of tables

- Table 1: Honeypot specification..... 19
- Table 2: Honeypot log format 23
- Table 3: Browser fingerprint 30
- Table 4: PDF Viewer JavaScript Support 54
- Table 5: E-Mail header fields..... 58
- Table 6: Additional E-Mail header fields 58
- Table 7: Honeypot results 66
- Table 8: Facebook test users..... 68
- Table 9: Facebook Test Case No Friend Suggestion..... 69
- Table 10: Facebook Test Case Friend Suggestions..... 70
- Table 11: Xing Test Case User ID Test 73
- Table 12: Xing Test Case Image Serach 76
- Table 13: Xing Test Case External Visitors..... 81
- Table 14: Test summary 90

1 Introduction

The need for privacy is deep-rooted in human beings. Privacy is hard to define, in different contexts, privacy has a lot of different definitions. [Unesco 01, page 9] The issue of privacy and personal data protection has grown in the last few years. Especially with the growing number of social network users and growing amount of personal information that can personally identify an individual. The obvious examples are somebody's name, address, date of birth, identification number, or photograph. Information that could be used for indirect identification is for example somebody's plate or credit card number.

A lot of personal information is provided on social network profiles like Facebook, on e-commerce platforms like amazon, or on career-oriented social networking sites like Xing or LinkedIn. They can be used for example for online advertisements or to put people under complete surveillance. Not only single data but also the combination of data is used to identify people online and to track their behavior. Online companies and organizations are using different tracking techniques and assign a person a unique identifier to be able to monitor the user's behavior. Such organizations do not need to know the real name of person. They do not need to know that the person's name is John Smith, but they know that the person's unique identification number is for example 1234567 and that he is interested in A, B, and C.

Are online users just a number? No, users are not just a number, the users could be identified by their assigned identification number but there is much more information and footprints that a user leaves on the internet. Nowadays a lot of organizations are tracking users online. There are a lot of purposes of tracking users. One of them is for example to improve the online shops. Another one is to know the user's behavior for different purposes.

The question in this thesis is, is it possible to reverse tracking? Is it possible to find out who is interested in private data about you and why is this person interested in this data? It would be helpful to know who visited your website, who visited different profiles or who searched for you with a search engine.

1.1 Problem

World Wide Web or abbreviated WWW is the most widespread and most popular service on the internet. WWW was established in 1989 by computer scientist Tim Berners-Lee. In the initial phase, WWW was simpler and clearer. There was not so much information about individuals.

Nowadays WWW is a very huge network. Information about people is stored on the hard drives of web servers all over the world. The WWW is growing and growing. It means there will be more and more information stored about individuals. With a growing amount of information on the internet about individuals, the potential for searching is increasing and the need for control of privacy is also growing.

In present times, the social networks are in widespread use and a lot of people visit them daily or even hourly. Many users are kids and teens, and some of them are criminals. No parent would like it if a criminal was looking for his kid or for his kid's photos on the internet. In the worst case, the criminal could be a pedophile. Therefore, it could be very helpful for parents to control who is looking and visiting social network profiles of their kids, who is looking for their addresses or who is mailing them, to take appropriate measures to prevent any potential harm.

Not only children but adults too can benefit from the information. For example, somebody gets information that a certain company was looking for him/her and downloaded his/her CV, it could mean that the company is interested in him/her and he/she should apply for a job. It improves the chance of finding a job.

Tracking the footprints and traces online users leave during the search for a certain person can answer the questions above or can deliver information to answer the questions.

The main problem is the size of World Wide Web in combination with restricted access to foreign data and anonymization methods for online surfing.

1.2 Objective

With the growing internet and growing amount of information about individuals all around the world, it is becoming increasingly important to know who collects, knows or searches for information about certain individuals. This thesis deals with the problem of tracking and identifying persons online who are interested in private data of a certain individual. Who is interested in information about this certain individual and the reason of interest?

Different methods for tracking a user online will be investigated and implemented in this thesis. If somebody visits your homepage and searches for information about you, it should be recorded. Then the records should be evaluated. On one hand tracking users online will be investigated. On the other hand, the opposite for tracking, reverse search methods will also be investigated and implemented. If somebody sends you an email with a request, to send your phone number or the CV, you might be able to find out more information about the sender than just what was presented in the email. You might be able to identify the sender.

The common services where private data is stored will be covered. Important parts of investigation in this thesis are social networks, homepages and searching engines. Some of them will probably deliver no results, but most of them might deliver very useful results.

The importance of investigation fields is calculated based on amount of information provided by a system, based on how often is a system used and based on the number of users in the year 2016. In the future, there will likely be new popular services and applications but this thesis's main task is to investigate the most common services in the year 2016. Some investigation examples are Facebook, Xing, LinkedIn and Google. Nowadays they provide the most information about individuals.

Find as much information as possible about people, machines or companies which are interested in you. The costs of implementing this thesis, including hardware and software, should remain near zero because every person should know more about his/her privacy without paying any money. It is possible that for different investigation fields (Facebook,

Google, File tracking) different software exists which can help but is not free. Only low cost variants will be investigated.

After investigating and implementing different methods, different algorithms, every method should be thoroughly tested. Test results should be investigated and an automatic mechanism should be developed if possible.

1.3 Construction

Firstly, it will be investigated why it is important to have more control and more information about private information redistributed on the WWW. Who is allowed to deal with this data and who is not? What kind of information do we provide to different services when we agree the general business terms? In addition to this, different limitations that could prevent positive results will be discussed but only theoretically. Are there any methods which are possible and which could be useful but their results are not useful, for example recording mouse moves on the homepage or identifying people based on keyboard typing. A very important part of this thesis will also be discussed, reverse search because in this thesis the main question is who is looking for me.

In the practical part, a lot of services will be investigated based on bugs, features and security holes which can provide us information needed to identify someone. At first, a honeypot will be created with many features for tracking a user online. HTML5 features will be used for some of the methods, for example for browser fingerprinting. An additional, the most common services such Facebook or Xing will be investigated each in an extra chapter. Their limitations will be also part of the practical part. An attempt will be made to develop a mechanism to make the practical part fully automatic or at least semi-automatic.

At the end of the thesis, implemented methods will be tested in real life. According to test results, a conclusion will be made. Which methods are more useful and why and which methods are less useful and why. What service does not give you any possibility to look into their log files and what service provides you the most information? These are the questions

that should be answered based on test results. The methods will be tested for their effectiveness.

2 Theoretical foundations

We live in a completely different world today thanks to the Internet, Google, Yahoo, Bing, Facebook and all other online platforms. No one's privacy is secure and no one's identity is safe. If you ever wondered "who is looking for you online", you are not alone. It is one of the most asked questions on Google.

Google processes about 1.2 trillion searches per year worldwide [Inet 01], figure 1 Google Searches shows how many searches there were per year in last few years. The number of searches is increasing and the amount of information about an individual is also increasing. For this reason, it is very easy nowadays to find someone on the internet. It is not only about finding a celebrity online, it is also about finding regular people. If you want to know something about your former school colleague, education of your boss or something about a person who borrowed money from you, you will probably look for him or her on the internet.

This applies to other people, but this also applies to you. If somebody is interested in you and in information about you, about your current home address, about your current appearance or about your car, he or she will try to find this information on the internet. It could be for example a job recruiter or it could be the bank or the money lender.

Year	Annual Number of Google Searches	Average Searches Per Day
2015	2,834,650,000,000	7,766,000,000
2014	2,095,100,000,000	5,740,000,000
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000
2010	1,324,670,000,000	3,627,000,000
2009	953,700,000,000	2,610,000,000
2008	637,200,000,000	1,745,000,000
2007	438,000,000,000	1,200,000,000
2000	22,000,000,000	60,000,000
1998	3,600,000 *Googles official first year	9,800

Figure 1: Google Searches

Nowadays there is no possibility to get a notification by any search engine if somebody was looking for your name. At this point, we encounter the next problem. If somebody is looking for "John Smith", how should Google know which "John Smith" is meant by searcher? Google should notify all John Smiths that somebody was looking for his name. The fact is that somebody has been looking for the name John Smith but you are not able to find out which John Smith is actually meant. In addition to this, if you have the same name as a very famous person, for example Jason Statham, the notifications would flood you. Finding out that somebody was looking for you becomes much easier with a unique first name and surname. If somebody is looking for your name and your name is unique, there are not so many options the searcher has and it is easier to track the searcher back.

Warren's and Brandeis' definition of privacy "*These considerations lead to the conclusion that the protection afforded to thoughts, sentiments, and emotions, expressed through the medium of writing or of the arts, so far as it consists in preventing publication, is merely an instance of the enforcement of the more general right of the individual to be let alone.*"(Warren and Brandeis 1890, p. 205), is a good early explanation of privacy when the internet has not yet existed. Information privacy exists when you are able to control the usage, release and circulation of personal information [Chung 01, page 1]. How can you control information about you on the WWW? At the moment, there is no possibility to be notified if someone was looking for your name, but you are able to control what private information will be provided about you. You are the person who provides the most information about you online, you are the person who uploads profile images to social network profiles, and you are the person who provides you email address or phone number online. If you want to know if somebody else provides information about you online, you can use Google Alerts [Wiki 01]. Google Alerts is a tool which is able to notify you if it finds new results about predefined terms by you, for example new web pages, newspaper, or articles containing the terms. You can add as many terms as you want. With this tool you can better control new data published online. If you control what data is published about you, you also control the surveillance area in which you have to make reverse searches to identify somebody who is looking for you. Smaller amount

of private data about you online means smaller surveillance area which means easier identification.

According to research from the year 2013 [PewRes 01], shown in the figure 2, users report that a wide range of their personal information is available online. About 66% of 792 interviewed adult internet users said that a photo of them is available online. 50% of them claim that their birth date is available online. As shown in the figure 2, a lot of additional personal information about them is available online. The challenge is to control circulation of this data and access to this data on internet.

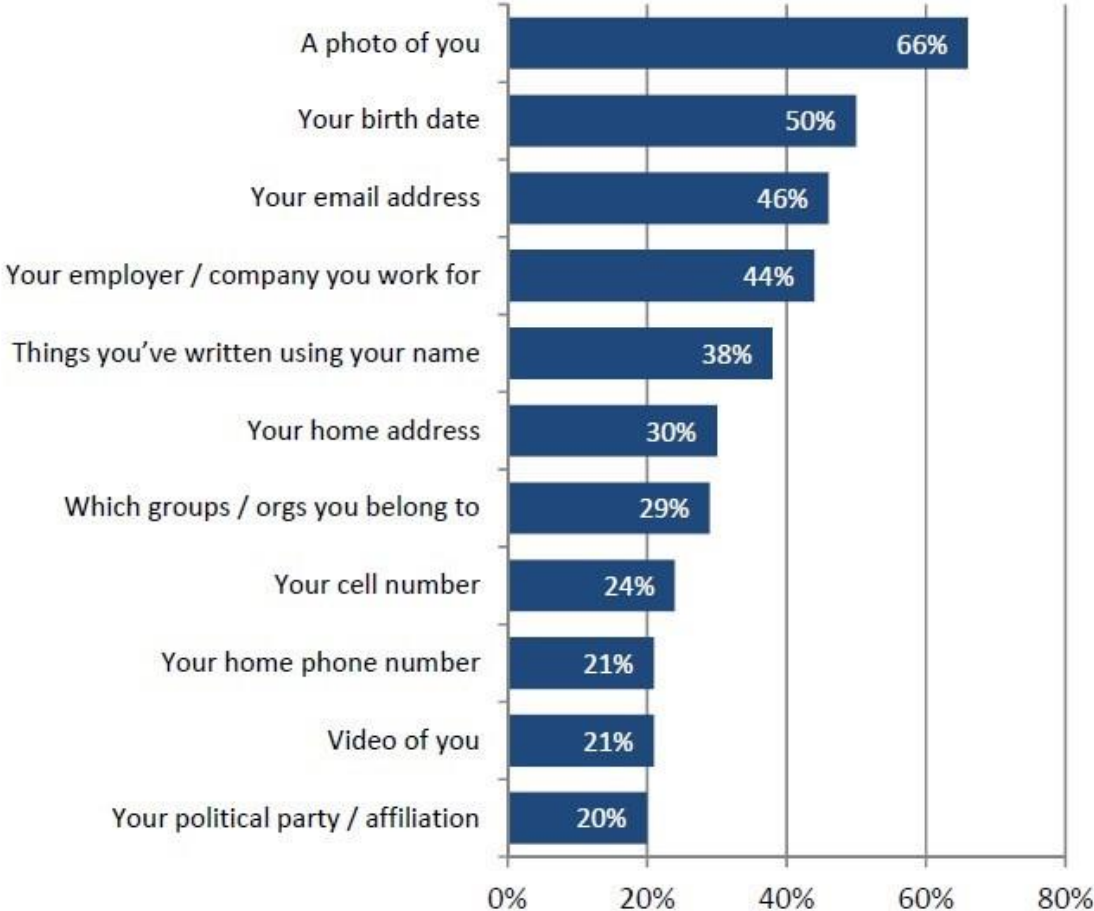


Figure 2: Personal information available online

2.1 Current ways to find out who is looking for you

Nowadays there are a lot of different ways to track people online. Most tracking tools deal with the improvement of online shopping. It means, these tools investigate only visitor frequency, time period where people shop the most and what products are most popular. These tools use cookies for tracking an online user. In this case, first party cookies will be used and the user is only observed by the site the user visits directly. There are also third party cookies which are typically hidden trackers such as networks embedded in most web pages. Third party cookies are able to obtain a user's browser history and a lot of other information through a combination of cookies and other tracking technologies. The best example for third party cookies is Google Adwords. The problem with this tracking mechanism is that the user is able to block such cookies or they are blocked by default, for example Safari. Additionally, these systems are not able to tell you if somebody is looking for you on platforms where you provide your private information, for example on Google or Facebook.

[OpenWPM_1_million_site_tracking_measurement.pdf]

Some organizations claim to be able to provide you information about who is looking for you online. Ziggs is for example one of them [Inet 02]. On Ziggs you can create a profile and market yourself. The profile will be indexed by Google and people will be able to find you. You get notifications if somebody visits your profile. The first problem with Ziggs is that all the information you get is predefined by Ziggs and you are not able to customize results to get more information or more useful information. The second problem is that Ziggs exists since July 10, 2013 no more. On July 10, according to <https://web.archive.org> the website www.ziggs.com was the last time online. There are some similar services to Ziggs but they are all working very similar to Ziggs. This tool has very good approaches. These ideas will help a lot in this thesis and they are used to develop a system that belongs to us, to develop a honeypot web page. Many helpful features will be implemented.

Similar features to Ziggs are provided by Google Analytics. Google Analytics [Google 01] is an online tool for tracking and reporting website traffic. Google launched this service in November 2005. Google Analytics is the most widely used web analytics software on the internet. There

are a lot of Google Analytics features for example Google Website Optimizer or Google Analytics e-commerce reporting. An advantage of Google Analytics against Ziggs is that you create your own webpage. You have to register this page on Google Analytics, then you are able to obtain statistics from Google Analytics. Some of Information you get from this tool are for example number of visits, country of the visitor or browser of the visitor. Doing analytics this way leads again to a problem with reliability of the system and dependency on Google. Why should Google or Ziggs do something for you if you can do it yourself? By doing it yourself, you retain control over your data. A much better solution is to create your own honeypot webpage and to have your own web server in your private network. In this case you have the ability to store only the information you need for identification. Additionally you are not dependent on other organizations. Combination of Google Analytics and logging on the honeypot webpage is possible but it brings no benefits. All information you get from Google Analytics, you get also from your honeypot webpage.

For the best known social network platform, Facebook, there are lot of existing fake tools and ways how to find who visited your Facebook profile, but no tool and no way works as promised. On this area there is a lot of undiscovered potential that will be investigated in this chapter.

Few third party services, for example Xing or LinkedIn, provide a feature to see all your visitors. These features will be individually short discussed or will be used in a combination with other self-implemented features to get more accurate results for easier identification.

2.2 Information needed for client's identification

Identification is a technology to identify specific objects from the traces they leave behind. The simplest and best known identification in real life is fingerprint recognition. Using fingerprints to identify individuals is a very old method of identification. Because of their uniqueness and consistency over time, fingerprints are very suitable for a successful identification. But before identifying a person using fingerprints, someone has to locate and collect usable fingerprints. The second step is fingerprint preprocessing for comparison. The last step is identification. For the identification, the fingerprint database is needed because the collected fingerprints have to be compared with known fingerprints. The more unique the fingerprint samples are in the database the easier it will be to identify someone of matching fingerprints. The fingerprint database should be up-to-date, meaning if there are fingerprint changes, the old fingerprint will be replaced by a new one. As shown in the real life example, the more information you have about a person the easier it will be to identify it.

This master thesis deals with the theme, identifying a specific client online. Is it possible to find as much information as possible to be able to identify someone online? To find out if this is possible, information to be found has to be determined. What is needed to identify someone online? To determine this information, first of all what we need to know is which tools will be used for searching for somebody.

The common way to be online is to use a web browser. The question is, is it possible to identify a web browser like fingerprint recognitions in real life? Theoretically it is possible, but practically the uniqueness of a web browser is easy to influence. This topic will be treated in a separate chapter. Following information about web browser will be collected:

- Public IP Address
- User agent
 - o Browser type
 - o Browser version
 - o Language settings

- Operation system
- Time zone
- Font
- Screen resolution
- Location
- Name
- Address
- Canvas fingerprint

From all collected data a unique fingerprint will be created for every web browser. Browser fingerprints will be used to get to know the user and to identify the user visiting a web site again.

A much better and much easier way to identify persons is to know their name, phone number, home address or to know what they look like, to have a photo of him or her.

2.3 Reverse search

Reverse search is a usual type of telephone inquiry. This technique allows you to find a name for a number. There are other reverse search techniques based on reverse phone lookup, too.

One of these techniques is reverse image search. This feature will be provided by some organizations. Some of them are Google with Google Image Search and TinEye with Reverse Image Search. These techniques allow you to upload a sample image and to search for a similar or same picture. This technique requires training models for every query. A lot of people assume that image search is conducted via fancy algorithms that determine the theme of the picture and index it. This assumption is wrong. There are a lot of characteristics of a picture that will influence the results.

The first important thing for image search engines is the file name. Second step is to look at the content surrounding it. Once an image search engine has crawled an image, looked at its

file name, and looked at its surrounding, it probably has an idea what the image is about. The next step is to find out which colors are used in the picture, and how frequently? Is there a face in the picture? What is the resolution? Based on it, the search engine finds a set of similar images. The algorithm is very complex, but these are the steps to be done for image searches.

On one hand image search is good in recognizing objects and landmarks. On the other hand this system is not so good for facial recognition. Nevertheless, this system is being tested in a separate chapter in context with XING or other platforms where you get an image of the visitor. In the figure 3, you can see one possible result for the query image.

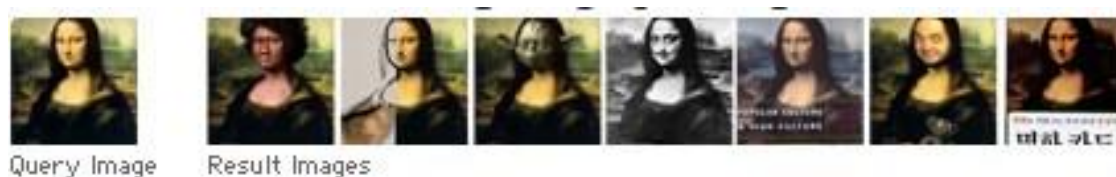


Figure 3: Image search example

A similar process to the telephone reverse search, finding a name for a number, is IP reverse search. In this thesis, the “phone number” of the visitor is his public IP version 4 address. If the user uses IP version 6, it would be his private and public IP address. Finding a person’s name and address to the given phone number is much easier than finding a person’s name and address to the given IP address. Phone numbers are more unique. IP address are not unique, it is also possible that more than one person works with the same public IP address. Additionally, public IP address of private users changes periodically. It is possible that the person A has public IP address A, but few days later, person B has the public IP address A. This can lead to confusion. This thesis is dealing with this problem: how to avoid confusion and how to make this information more unique. What information can you find to the given IP address is also a question which is helpful for identification.

3 Practical Part

3.1 Honeypots

Many different terms, definitions and classification for honeypots exist. The first concept of “honeypot” has been introduced in the late 80's. In “stalking the Wily hacker“, Clifford Stoll describes the first concept of honeypot in 1986. A computer intruder attacked the Lawrence Berkeley Laboratory (LBL). They decided to allow the intruder access to their network while they made records of his activities and they retrace them to his source. They wanted to find out who was breaking into their system. To do this, they allowed the intruder to use their computer to reach many others. This was helpful to find out how the intruder was breaking into their system. Additionally they wanted to find out the weaknesses of their system based on documentation made during the breaking-ins. It was the first concept of today's honeypots.

The “Honeypot” definition which could be found in many articles and many other articles refers to it, is from book “Honeypots, tracking hackers” written by Lance Spitzner, a senior security architect for Sun Microsystems says:

“A honeypot is a security resource whose value lies in being probed, attacked or compromised.”

[Spitzner 01, page 40]

In further steps, Lance Spitzner explains what it means. It means, whatever you designate as a honeypot, the goal is that your system should be probed, attacked and should possess weaknesses. It does not matter what the honeypot actually is (router, web server, and different services running on a virtual or non-virtual machine). But what is important? It is important that the potential attacker is convinced that he broke into a real system and got some important information.

What information will be provided in our context on a honeypot? All information that has no production value will be provided in a honeypot, it means information about a person that you want to use as bait for visitors.

This point is the most important feature of honeypot for this master thesis. With a honeypot you are able to provide information about yourself that you want to provide and at the same time you are able to record behavior of visitors.

The main goal of honeypots in this thesis is to attract people to a honeypot website you provide to record their behavior to find out why someone searches exactly the information provided on a honeypot. It means, in this thesis the honeypot resource will be a web server, consisting of a computer with a HTTP (Hypertext Transfer Protocol) server (software).

Honeypots are usually located separated from the rest of the productive network. There are few possible locations for the honeypots, Figure 4:

- outside of a firewall
- inside of a firewall
 - DMZ (demilitarized zone)
 - internal Network
 - VLAN (virtual LAN) for the honeypot
- Cloud Based Webserver

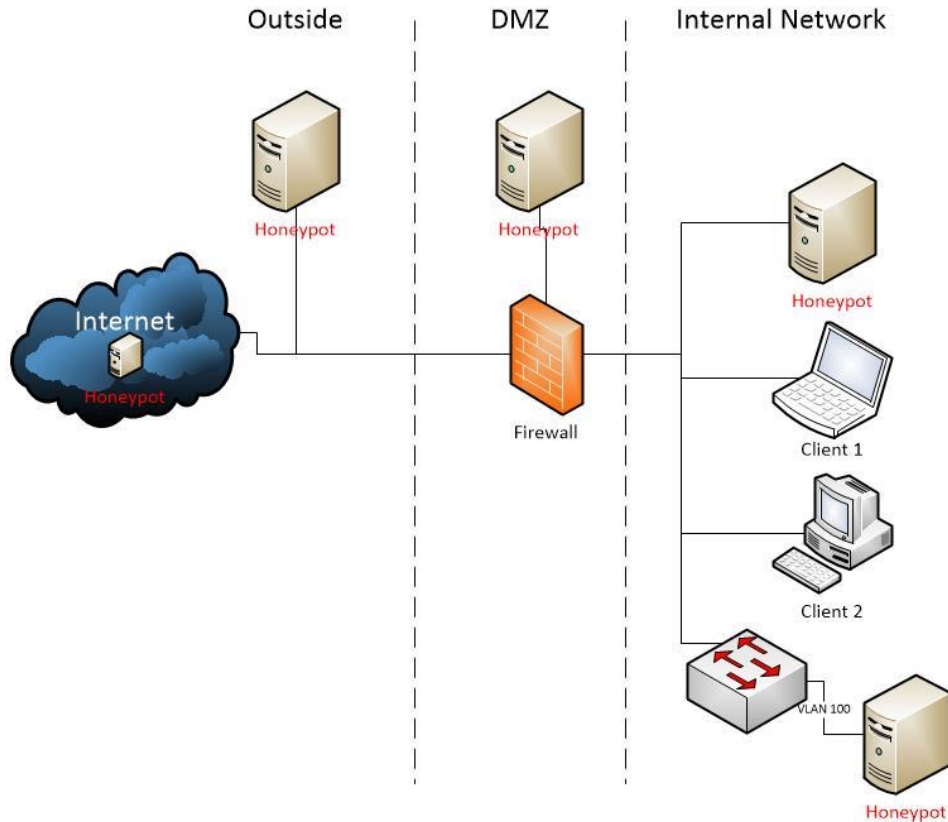


Figure 4: Honeypot location

In this thesis there is no need to put the honeypot into the production network. In fact, it is more secure to locate honeypots separately from the clients and servers as shown on the right in the figure 4, called internal network, so they can be excluded. The decision should be made between the outside location, left part in figure 4, and DMZ in the middle of figure 4 or maybe a solution with a special VLAN for honeypots. The outside location is not suitable for home networks, the reason is that the ISP (internet service provider) modem or router is also at the same time representing the firewall. Therefore, you do not have access to devices outside of your ISP device and it is impossible to control a device outside of your firewall. The remaining possibilities are DMZ or a special VLAN in the internal network. For both solutions a suitable device is needed. On the one hand, for the DMZ solution, the ISP device has to provide the possibility to locate one or more devices in the DMZ. On the other hand, for the special VLAN solution, an additional switch that provides VLANs will be needed in the internal network. The last solution is a cloud based server. In this case you put more data online and the surveillance area becomes larger because all your webserver data is stored somewhere in the cloud.

Another disadvantage of a cloud based webserver are flexibility of the system and dependency on the provider. The biggest disadvantage in context of this thesis is the necessary adoption of own processes to the used cloud system.

In this thesis, the honeypot server will be placed in the DMZ because the ISP router provides the DMZ feature and no additional equipment is needed.

3.1.1 Honeypot specification

There are no guidelines on how to build a honeypot properly. There are no rules on what a honeypot should look like or what a honeypot should be. Everybody builds their own honeypot according to their needs. A lot of existing honeypots could be used as a reference for building a new one. In this thesis, a new honeypot will be built and prepared for our needs.

The honeypot system is built from the ground up as a customized honeypot. There are two general categories of honeypots, production honeypots and research honeypots [Spitzner 01, page 43]. The concept of these categories comes from Marty Roesch, developer of Snort. Production honeypots protect an organization, while research honeypots are used to learn. Research honeypots are designed to learn about attackers or visitors, who they are, how they are organized, what kind of tools they use to attack or visit the honeypots. Because the main objective of this work is to find out “who is looking for me online”, a kind of research honeypot will be designed and implemented in this thesis.

3.1.1.1 Hardware

To be used efficiently, all computer software needs certain hardware or other software to be present on a computer. The system requirements for our honeypot are not high. As shown in the table 1, our honeypot is running on a common office PC. For sure, a Raspberry Pi could be also used for honeypot because of its low profile, minimal power consumption and its price, but it is more complicated to use than a common desktop PC. This also can be done with a

virtual environment but for this purpose the host machine has to have enough resources. Honeypot hardware needs to be fast enough to run the operating system and web server you wish to install.

One advantage of a honeypot built in this thesis is that it collects very little data, but the importance of data it does collect is very high. Instead of logging gigabytes of data every day, the honeypot will only collect less than 1 megabyte of data every day. Another important advantage of honeypots is the resources. A honeypot can work with little resources. For example, a firewall may fail because its connection table is full, so the firewall blocks all connections instead of blocking only unauthorized connections. On the other side, Intrusion Detection Systems (IDS) that monitors a network or system for malicious activity could have too many network activities to monitor. When this happens, the IDS's sensor becomes full and the IDS doesn't work. With a honeypot you do not have such problems.

3.1.1.2 Software

Only one additional software will be installed on the Honeypot system. It is XAMP v3.2.2. The purpose of XAMP is to act as a web server and to log everything happening on the honeypot webpage. Logging is the most important part of this chapter. Additional software for visualizing or for evaluation will be installed on another machine.

3.1.1.3 Router

Depending on the kind of honeypot which will be used, different ports have to be opened and forwarded on the firewall and ISP (internet service provider) router. Some honeypots need a lot of open ports because they can emulate many services. The Honeypot built in this thesis will only need two ports for two different services because we only want to find out who is looking for us on the World Wide Web and we only have a web server running on it. Therefore, two services which are open on the firewall and forwarded by the router to the honeypot PC are TCP (Transmission Control Protocol) port 80 for http and TCP port 443 for secure http, it means https.

Additional settings have to be made on the router. If you use dynamic DNS you have to configure it also on the router. Dynamic DNS allows you to configure your web server with a non-static public IP address accessible from everywhere without knowing your current IP address. Dynamic DNS makes automatically updates between public IP address and DNS. Additional, you have to create an account with the DDNS service in order to use DDNS. More details are in following chapters.

Location	DMZ
Category	Research honeypot
Resource	Web Server
Hardware	Personal Computer CPU : Intel core i5-2500 @ 3.3 GHz RAM: 4.00 GB HDD: 500 GB
Software	Windows 7 Professional Service Pack 1 HTTP Server: Apache (XAMPP v3.2.2)
Description	Static private IP address Dynamic public IP address LAN Connection

Table 1: Honeypot specification

3.1.2 Creating the honeypot website

To get online users to visit a web page with the hope to find what they are looking for, the honeypot web page should appear like a real web page, not like a test web page or web page without content. The content of a web page is very important. Nobody will visit the page if there is nothing to read or nothing to download. Therefore, the content of a honeypot web page is crucial.

The web page in this project will look like a personal web page with a lot of information about Dusan Jaguzovic, the creator of this project and author of this thesis. A brief text about Dusan Jaguzovic is provided on the web page. On the web page navigation there are also hobby, images and contact items. There is a lot of dummy information about the creator, some information is intended to be read and some to be downloaded. Why present dummy information and not real information about a person? If you want to catch a mouse, you will probably not use a whole refrigerator with fresh food to catch it. You will probably use a piece of old cheese because the losing of small piece of cheese does not hurt like losing a whole refrigerator.

For tracking purposes we have prepared two PDF documents for download:

- CV_Dusan_Jaguzovic.pdf
- Address_Dusan_Jaguzovic.pdf

The same documents are available for download in DOCX format:

- CV_Dusan_Jaguzovic.docx
- Address_Dusan_Jaguzovic.docx

The use of PDF and DOCX files will be explained in the chapter File tracking.

As you can see in the figure 5 Appearance, the design does not look suspicious.

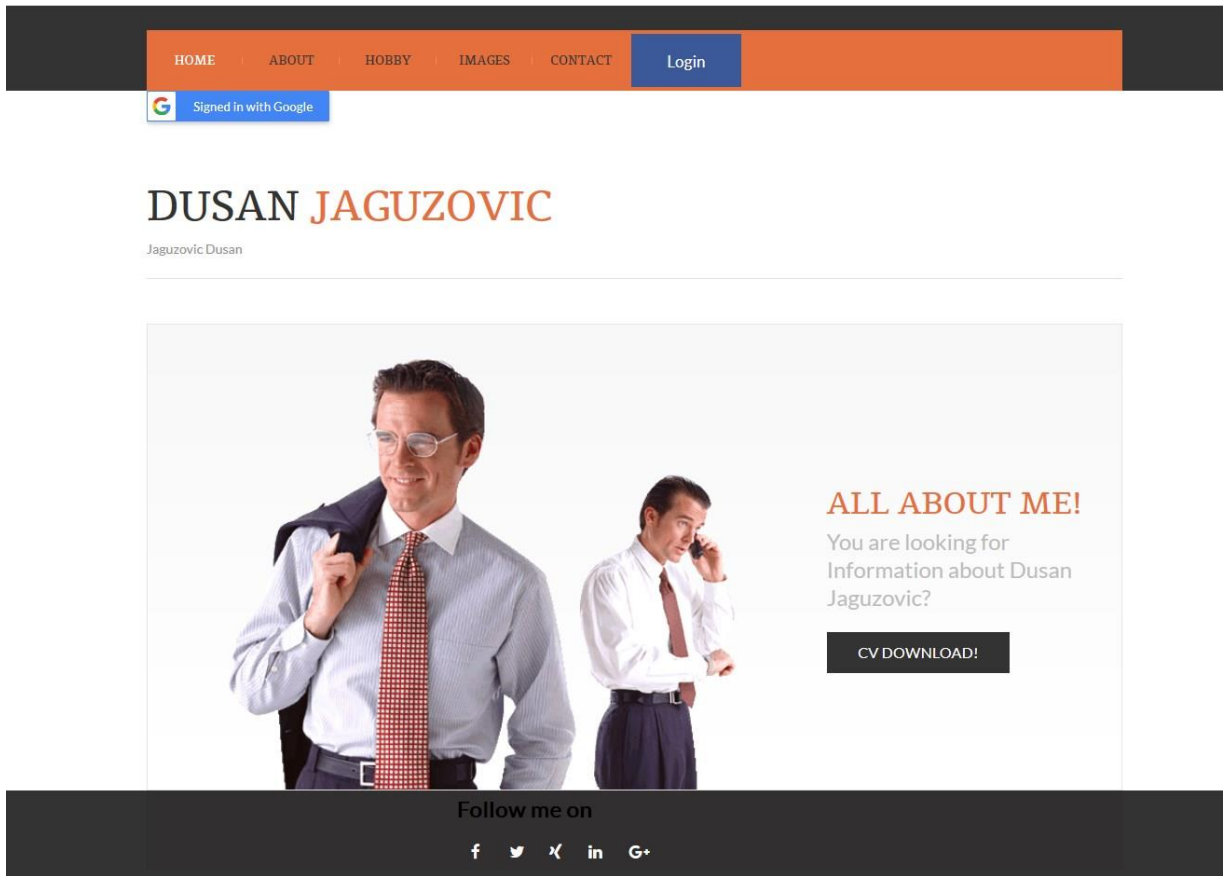


Figure 5: Appearance

Another real productive web page with real information about you should not be a problem. If somebody is looking for information about you and both web pages are on the first result page after Google search, this person will probably visit both of them if he is really interested in you. The best case for identification is when the honeypot webpage is on the first result page and the real one somewhere else. The worst case is when the real web page is on the first result page and honeypot web page on another page. Therefore, search engine ranking is very important. More details about web page performance and reachability in a chapter number 3.1.3.

3.1.2.1 Honeypot features

3.1.2.1.1 Logging

“A computer file that contains a record of all actions that have been done on a computer, a website, etc.” is the explanation of the word “log file” of Cambridge English Dictionary. Logging is the act of keeping the log files. In this thesis, logging is very important.

Log files on the honeypot server are of great importance for the evaluation. A HTTP Server that is installed on the honeypot machine provides different mechanisms for logging. There are few different log files:

- Access.log
- Error.log
- Php_error.log

Access.log file is the most important log file in this thesis for our purpose. This log file contains all requests processed by the server. The location and content of the log file are not predefined. The log format will be used to predefine how the server records information in the access log. What information is needed and what information make sense for logging. For our purpose, custom log format will be used, shown in the table 2. Custom log format will be created because it gives you ability to log what you want and what you need.

LogFormat "%h %l %u %t \"%r\" %>s \"%{Referer}i\" \"%{User-Agent}i\""	
Key	Explanation
%h	IP address of the client
%l	Remote logname if available
%u	User ID of the person requesting the document
%t	The time that the request was received

<code>\">%r\</code>	First parameter is method used by client. Second parameter is the requested resource. Last parameter is the protocol used by client.
<code>%>s</code>	HTTP Status code
<code>\"%{Referer}i\</code>	This gives the site that the client reports having been referred from.
<code>\"%{User-Agent}i\</code>	This is the identifying information that the client browser reports about itself.
Log file entry example	
<pre>213.162.68.98 - - [03/Oct/2016:16:45:40 +0200] "GET /images/img-1.png HTTP/1.1" 403 "http://dusanjaguzovic.ddnss.de/" "Mozilla/5.0 (Linux; Android 6.0.1; SAMSUNG SM- N910F/XXS1DPH7 Build/MMB29M) AppleWebKit/537.36 (KHTML, like Gecko) SamsungBrowser/4.0 Chrome/44.0.2403.133 Mobile Safari/537.36"</pre>	

Table 2: Honeypot log format

The log file entry example shown in the table above contains a lot of useful information. In addition to already explained records there is a record within the User-Agent information that is very useful and helpful for evaluation. We can see the entry “Android 6.0.1; SAMSUNG SM-N910F/XXS1DPH7 Build/MMB29M”. If we analyze the record we will see that the request was sent by a smartphone Samsung N910F, Samsung Galaxy Note 4. In addition to this information we get is the build number of the device. This is very helpful information that will have great importance in the evaluation of recurring visitors.

Every time a user visits the honeypot web site, a log file entry is made. Every movement on the web site will be recorded, it means if the user visits the start page, then downloads the PDF file and then visits the contact page, everything will be logged.

Storing the information in the access log file is only the start of log management. The next step is to analyze information recorded and stored in the log files. Analyzing the log files is not part of the job of the web server. Information is stored and at the end it will be combined with information obtained using other honeypot features. Using the relationship between few data sets of results could identify a person even if a single set is not able to.

3.1.2.1.2 Geolocation

Geolocation feature as the name tells is a feature that identifies the real-world geographic location of an object. In this thesis it will be the location of the searching client. Geolocation is a very important feature in this thesis. If it works fine, it means the accuracy is good.

Latitude and Longitude

Depending on the method used for geolocation, a geolocation can be as general as the continent or country the user is connecting from or it can be his geographic position, defined as latitude and longitude coordinates. Latitude and longitude coordinates can be converted to a specific place or specific address. The second method is more complicated and requires user permissions because of data privacy. It means, you are not able to locate the user's latitude and longitude coordinates without the user's permission. You have to ask the user for permission to locate the client. Additionally, it requires the web browser feature for geolocation; the device has to support Global Positioning System or GPS. The results are very accurate. The deviation resulting from a few tests is plus or minus 20 meters. This information is very useful for identifying clients visiting your website. The JavaScript code below Code XY shows the implementation of geolocation with permissions.

```
function getLocation() {
    if (navigator.geolocation) {
        navigator.geolocation.getCurrentPosition(showPosition);
        //Geolocation is supported by this browser
    }
    else{
        //Geolocation is not supported by this browser
    }
}

function showPosition(position) {
    var x = position.coords.lattitud;
    var y = position.coords.longitude;
    //Write x and y to logfile
}
```

The result for variables x and y from the source code shown above Code XY are latitude and longitude coordinates. There are many services provided on the internet that convert such

coordinates to a specific address. Getting reverse geocoded address from latitude and longitude is a big step in identifying somebody. The result in the form of latitude and longitude will be stored in a log file.

The World Wide Web Consortium or W3C stipulates that the location determination may only be carried out with the express consent of the user. The user will be asked for permission to determine his location. In this case, you should “persuade” the user to accept the determination of his location. One possibility to do this is to promise the user more detailed information about you if he accepts. Another possibility is to promise the user that the web page will be translated into the language of the country from where he actually visited your web page. After he accepts the GPS determining of location, you are able to get the longitude and latitude coordinates.

Limitations with latitude and longitude coordinates are devices that do not support GPS, for example desktop PC or some other mobile devices. In this case this feature is not useful. No entry will be made in the log file.

Location based on IP address

Another way to locate someone is based on their IP address. It is true that your IP address does not reveal your address or your identity, but it can reveal what city or even general area you are in. Location based on IP address is not as accurate as the first method described in this chapter, location based on latitude and longitude coordinates. An advantage of this method is that you do not need any additional permissions or additional web browser features. Locating will only be done with the help of the IP address.

How is it implemented? There is a predefined API for this geolocation method available on the internet called ipinfo.io. You send the public IP address you want to localize and you get a one-dimensional array as a result. The result array contains the following information (if available):

1. Public IP Address
2. City

3. Region
4. Country
5. Location (latitude and longitude coordinates)
6. Postcode

In addition to this information, the time stamp will be stored,too. For checking the visitors IP address internet connectivity is required because geolocation of the given IP address is queried online. It is more accurate because database updates are made automatically.

In contrast to the first method, accuracy is not so high. The accuracy resulting from few tests deviates from country to postal. Sometimes it is possible to get all information shown above from a public IP address. Sometimes it is only possible to get the country and latitude and longitude coordinates of the capital of the county. In the log entry shown in Figure GeoLocationEntry you can see two different geolocation results based on IP address. The first one with the IP address 77.117.2.95 was located in Austria, Vienna (based on latitude and longitude coordinates). The second one is more accurate. Geolocation was able to find all information for the IP address 90.146.185.145.

```
77.117.2.95; ; ; AT; 48.2000,16.3667; ; 17-12-2016; 02-42-58 PM  
90.146.185.145; Ansfelden; Upper Austria; AT; 48.2097,14.2900;  
4052; 17-12-2016; 07-49-48 PM
```

Finding somebody's address and its geographical coordinates is highly relevant for identification. It is one of most important information you get. If you know the visitor's address, the probability to find this person is very high. The limitation with this method are dynamic public IP addresses. It is possible that more different clients has the same public IP address, in the same city, if they have the same internet provider but not at the same time. It means if you get geolocation of a person at certain time, it is possible that another person will have the same IP address few days later. If you want to identify someone, you should do this at the time of visit.

More accurate results would be possible through announcing the database or parts of database of geolocation positions of a lot of wireless routers collected by Google. Google

StreetView cars does not only collect the GPS data but also a lot of wireless data. The knowledge about the location of these wireless routers (MAC address, SSID, and coordinates) allows the calculation of client's location. Not only Google StreetView cars collect these information but also your own mobile device. Google checks periodically your location using GPS, Cell-ID and Wi-Fi information. In this case your mobile device will send back the SSID and MAC data. Knowing this information would improve the identification results. There are few free Wi-Fi databases that provides you WLANs and their positions. These databases are not as big as Google databases and they are not trustworthy since everyone can register information he wants.

3.1.2.1.3 Browser fingerprinting

Browser fingerprinting is a technique that is used on a website that can generate an ID or fingerprint that is unique to your browser. Its purpose is to make your browser trackable. Browser fingerprinting is primarily based on browser, operating system and installed plugins. It does not make a person working on a computer trackable but the browser installed on the machine. Browser fingerprinting is a technique with great accuracy. According to panoptclick, research project of the EFF (Electronic Frontier Foundation), [EFF 01, Page 2], 83.6% of browsers have a unique fingerprint.

Some people would say that cookies do the same thing as browser fingerprinting. Browser fingerprints and cookies are completely different techniques. There are no cookies required for fingerprinting, but in this thesis they will be only used to detect changes of browser fingerprints easier. If the fingerprint changed and the cookie did not change, it means that there are some system changes on the visitor side. There is also no browser option "Do not track me" that can make a difference as with cookies. Deactivating flash or deactivating java in your browser would not prevent you from tracking. Deactivating JavaScript makes the tracking for someone a little bit harder but nowadays almost every website requires JavaScript. JavaScript is a platform that allows Facebook, Google Maps or Gmail to function. You are able to delete or deactivate cookies but you are not able to avoid browser fingerprinting. Client-

side scripting languages allow the website developer to collect a lot of information like screen resolution, plugins, operating system or browser type without user's knowledge.

The advantage of browser fingerprinting is that the extraction of browser fingerprints is happening automatically in the background without the user's knowledge, it means fingerprints do not require user's permission. Browser fingerprinting algorithm collects a number of commonly and less-commonly known characteristics that the browser makes available to the website.

In this project, a browser fingerprint library developed by Valentin Vasilyev, web developer from Russia, will be used. The library is MIT licensed.

Key	Possible value	Explanation
User agent	"Mozilla/5.0 (Windows NT 10.0; WOW64; rv:50.0) Gecko/20100101 Firefox/50.0"	HTTP header sent to the server that contains information regarding your browser and operating system
Language	"de"	HTTP header sent to the server that indicates the preferred languages for the response
Color depth	"24"	Browser-populated JavaScript attributes that indicate the resolution of the device's screen (window.screen.colorDepth)
Pixel ratio	"1"	
Resolution	"array[2]; array[0] = 1920, array[1] = 1080;"	Browser-populated JavaScript attributes that indicate the resolution of the device's screen (window.screen.height/width)
Available resolution	"array[2]; array[0] = 1920, array[1] = 1040;"	Browser-populated JavaScript attributes that indicate the resolution of the device's screen minus 40px for toolbar on windows 7/8/10
Time zone offset	"-60"	Timezone offset of your browser obtainable through JavaScript (new Date().getTimezoneOffset())
Session storage supported	"1"	JavaScript test to find out if session storage is supported (storage of a specific value in "sessionStorage")
Local storage supported	"1"	JavaScript test to find out if local storage is supported (storage of a specific value in "localStorage")
Indexed DB	"1"	Indexed database available or not

Key	Possible value	Explanation
Open database	"1"	
CPU Class	"unknown"	Returns the central processing unit (CPU) class of the user's operating system.
Navigator platform	"Win32"	Browser-populated JavaScript attribute that indicates the platform the browser is running on (window.navigator.platform)
DoNotTrack	"unspecified"	Browser-populated JavaScript attribute that indicates your Do Not Track setting (window.navigator.doNotTrack)
Regular plugins	"array[3]; array[0] = ActiveTouch General Plugin Container::ActiveTouch General Plugin Container Version 105::application/x-atgpc-plugin~gpc, array[1] = Citrix Receiver::Citrix Receiver Plugin (Win32)::application/x-ica~ica, array[2] = DYMO Label Framework::DYMO Label Framework Plugin::application/x-dymolabel~"	Browser-populated JavaScript attribute that gives the list of activated plugins in the browser (window.navigator.plugins)
Canvas fingerprint	"canvas winding:yes~canvas fp:data:image/png;base64,iVBORw0K GgoAAAANSUheUgAAB9AAAADICAYA AACwGnoBAAAgAEIEQVR4nO3db4xd5 Z3g+R8siVeh06lovMnSk0AliAWR9IBD ETtNa0RIV8nsv2mVNJAZotbKypoLzg5 Srdz9YjbTgm6topXW07XavGgVmR7vz MRpqdMj04iROukmhQTBXj~~~~~ ~/d1AAAAAAAgCtUO5wv9hvND8ajF jKiKW8P57KiKU8HGfHUpx9mMWO p8fVMczgNxtO9wvhD9RPMnMuKp/m L6dDumr/S4bGMp4zsnBv01181Kv7+x L+9pHKg9aG/luKnx1YzG0sGxPMMnd/ 9Gntpzlk/tyeqx+3t5evbGcVwLAAAAA AAAqFtj6eloLGVnPPRHvxkPPPbNUU6 Zq8u09x3O!8B9TzyxJqQ/Gz8aazwvAv o98URGLG0a0Fei33Ce7XC+1H9AL4/5z DhUuvQ/PjxsQF8b0htLB6OxdNUF9HH NQM/Ts"	Rendering of a specific picture with the HTML5 Canvas element following a fixed set of instructions. The picture presents some slight noticeable variations depending on the OS and the browser used.

Key	Possible value	Explanation
WebGL fingerprint	"data:image/png;base64,iVBORw0KGgoAAAANSUgAAASwAAACWCAYAABkW7XSAAARckIEQVR4nO3c/2vbi37f8eefsR82uD8cdtj5ITTQQGhgEYGZGfyDmWEepoYZ/IOpN8NMDTWY2cPUMA+3hmqYYeqLGWaGupga5hJKUEYogVyak+Q4dfwFWYqQriJFU6ToSihS9NwP95Z2cO/p+ZLEtvJ~~~~~shader high int precision:0~webgl fragment shader high int precision rangeMin:31~webgl fragment shader high int precision rangeMax:30~webgl fragment shader low int precision:0~webgl fragment shader low int precision rangeMin:31~webgl fragment shader low int precision rangeMax:30"	Rendering of specific 3D forms following a fixed set of instructions. The picture presents some slight noticeable variations depending on the device of the user.
Is Adblock installed	"false"	Test to find out if the Adblock extension is installed
Has lied languages	"false"	Overwritten the browser identifier language
Has lied resolution	"false"	Overwritten the browser identifier resolution
Has lied operation system	"false"	Overwritten the browser identifier operation system
Has lied browser	"false"	Overwritten the browser identifier user agent
Touch screen detection and capabilities	"array[3]; array[0] = 0, array[1] = false, array[2] = false"	
JS Fonts	"array[3]; array[0] = "Arial", array[2] = "Calibri", array[3] = Century;"	Flash attribute that gives the entire list of fonts installed on the operating system (flash.text.Font.enumerateFonts(true))

Table 3: Browser fingerprint

As shown in the table above there are 24 characteristics that are available through the browser to the fingerprinting algorithm. Few of them are self-explanatory. Some of them are a little bit more complicated to understand but still very important for building the fingerprint.

Canvas fingerprinting is a browser fingerprinting techniques that allow websites to create fingerprints using HTML5 canvas element. This technique is similar to common browser fingerprinting but the difference is that canvas fingerprinting gets the information through the canvas API of modern browsers. Browser generates a hidden graphical HTML5-Image element and renders it on the web page during the visit. The appearance of the hidden graphic is different for most browsers (caused by CPU, graphics driver, browser, etc.). Exactly the difference between appearances of the graphic makes the fingerprint unique. In the figure 6 you can see the superimposed images that should show the difference between different browsers.



Figure 6: Canvas fingerprint

A lot of well-known web sites use canvas fingerprinting algorithm to track and identify their visitors, for example <http://t-online.de> or <http://wetter.com> . The result of canvas fingerprint is a hash value as you can see in the table fingerprint sources.

Canvas Fingerprint Process:

1. A hidden canvas element is created
2. Shapes are drawn on the canvas
3. Text is written on the canvas
4. Canvas data is converted from binary form into Base64 encoded string
5. This string gets appended to other components
6. All components are passed through a hash function -> result

WebGL is similar algorithm to canvas fingerprinting. It works with image rendering and gets a hash value as result, you can see in the table fingerprint sources.

The result of browser fingerprinting is a unique ID like “c6c276cbd38d362fca8cbba7c4aaf774”. This unique ID stays the same for the same browser without new extensions, without new installed updates, with the same hardware for example monitor. Even if there have been minimal changes the fingerprint will not be the same. This is one of the biggest problems for fingerprinting.

Browser fingerprint will be saved in the database in combination with other information we get with other honeypot features. In addition to fingerprint, there will also be the public IP address of the visitor and the timestamp stored. It should help to track the user. If one user always uses the same browser and the same device to be online, the fingerprint will be always the same. Public IP address might change caused by the geographical movement of the smartphone and his owner or by non-static IP address of ISP router. On the other hand, if the public IP address does not change and browser fingerprinting algorithm gets different browser fingerprint for the same IP, it could mean that the user made software or hardware changes on his system or more than one user works behind the same public IP address, for example a company with one public IP address. Only for this purposes, cookies are also stored. Cookies should only help to distinguish between making a change of system and being different person. If the fingerprint changes, the public IP address is the same and the cookie is the same, means

that the client made some changes on its system. If the cookie and the fingerprint are not the same but the public IP address is the same, the probability that it is a different person is very high. It is also possible that the user made system changes and cookie expired or is changed, but the probability for this scenario is very low. The expiration date for the cookies set is very long.

There are also problems with browser fingerprinting. They are not always reliable. You can also thwart the browser fingerprinting. Some of possibilities are:

- Using the Tor browser
- Blocking JavaScript from loading in the browser
- Using NoScript browsers extensions to block JavaScript from known fingerprinters

3.1.2.1.4 APIs

An API, or application programming interface, enable software programs or internet sites to communicate with one another. There is a set of predefined methods of communication between software components. There are different APIs, there are APIs for web-based systems, for operating systems, for software libraries etc.

If you want to find out who is visiting your web page with help of different APIs, you should act as the man in the middle, see figure 11. You should build a kind of bridge between your website and the application that should help you at identifying this person.

The most popular APIs are Facebook APIs. There are different ways to implement Facebook APIs on your homepage. You are able to implement Facebook Like and Share Buttons. These buttons provide the user the possibility to share content of your homepage on its Facebook profile. Another useful Facebook feature is to implement Facebook Login on your homepage. Facebook Login on your homepage enables people to sign in into your web page with their Facebook credentials. In this case, you are able to identify the visitor of your honeypot webpage by its Facebook ID that you are able to get if somebody logged in with Facebook credentials. To implement Facebook Login on your homepage, you need a Facebook App ID before you start. In this case, dummy Facebook App ID is created and is used for the Facebook

Login. The name of the Facebook App is also called “Dusan Jaguzovic”. This should arouse visitor’s interest for “Dusan Jaguzovic”.

Facebook Login request is made automatically when somebody visits the honeypot website. This feature is only implemented on the homepage. Users are asked to login with Facebook credentials.

You can download the source code and documentation for the Facebook API on the <https://developers.facebook.com/docs/facebook-login/web>. The most important information you get with Facebook API if the user logs in on your honeypot webpage is the user’s Facebook ID. In most cases, adding user’s ID <UID> in following links will show you the profile of the user. Example for Facebook ID is 10152384781676191. In some cases you are able to see the user’s profile only if you are logged in. It depends on profile settings and privacy settings.

```
http://facebook.com/profile.php?id=<UID>  
https://facebook.com/<UID>  
https://www.facebook.com/app_scoped_user_id/<UID>
```

You can download the source code example with documentation for Google API on the <http://www.codeworld.com/login-with-google-account-using-javascript/> . In this case you can get the Google Name of logged in user.

3.1.3 Honeypot availability

Web page reachability is an important factor. It plays a key role in crawling, indexing and ranking of web pages. Web crawlers and search engines measure website performance in terms of page speed, navigability or reachability. Additionally it plays a role in finding your web page by users. For this reasons, the web page reachability should be as high as possible.

Making web page reachable from consist of few important steps.

1. Guarantee IP reachability of webpage (IPv4 and/or IPv6)
2. Guarantee DNS reachability of webpage (dynamic DNS)

At first, the webpage should be available from outside by internet protocol IP. Nowadays IP version 4 (IPv4) is the most common networking method in the internet. IPv4 uses 32-bit addresses. The first problem with IPv4 addresses is that these addresses are not unique, it means there is possibility that two or more different clients visit the honeypot web page with the same IPv4 address, with the translated IPv4 address by ISP router. This leads to problems with identification. Duplicate IPv4 addresses make identification more difficult. The second problem is that internet is running out of unused IPv4 addresses. IPv6 is designed to solve the address shortage and to eventually replace IPv4. IPv6 uses 128-bit addresses, theoretically allowing 2^{128} . Over 300 addresses are possible per square meter of the earth. It makes IPv6 unique. Every IPv6 client is unique and this helps in identification. Nowadays making webpage reachable via IPv6 is not the most popular way. Most webpages are only reachable via IPv4. On the other hand there are also limitations in relation to IPv6. According to Google, maximum 16% of users worldwide access Google over IPv6, Figure 7 IPv6 Google. Additionally, if you look at the chart that shows the availability of IPv6 connectivity in Europe, figure 8 IPv6 Europe, you see the low IPv6 availability. IPv6 availability in Austria is 4.87%. [Google 02]

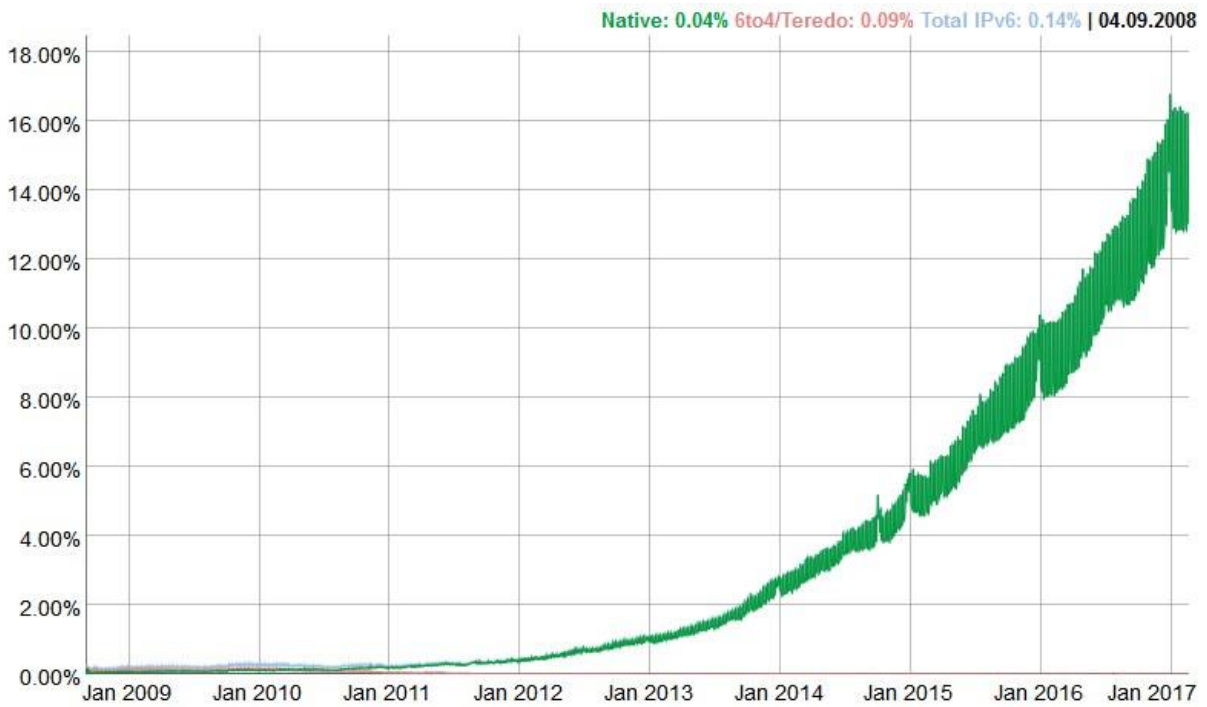


Figure 7: IPv6 google access

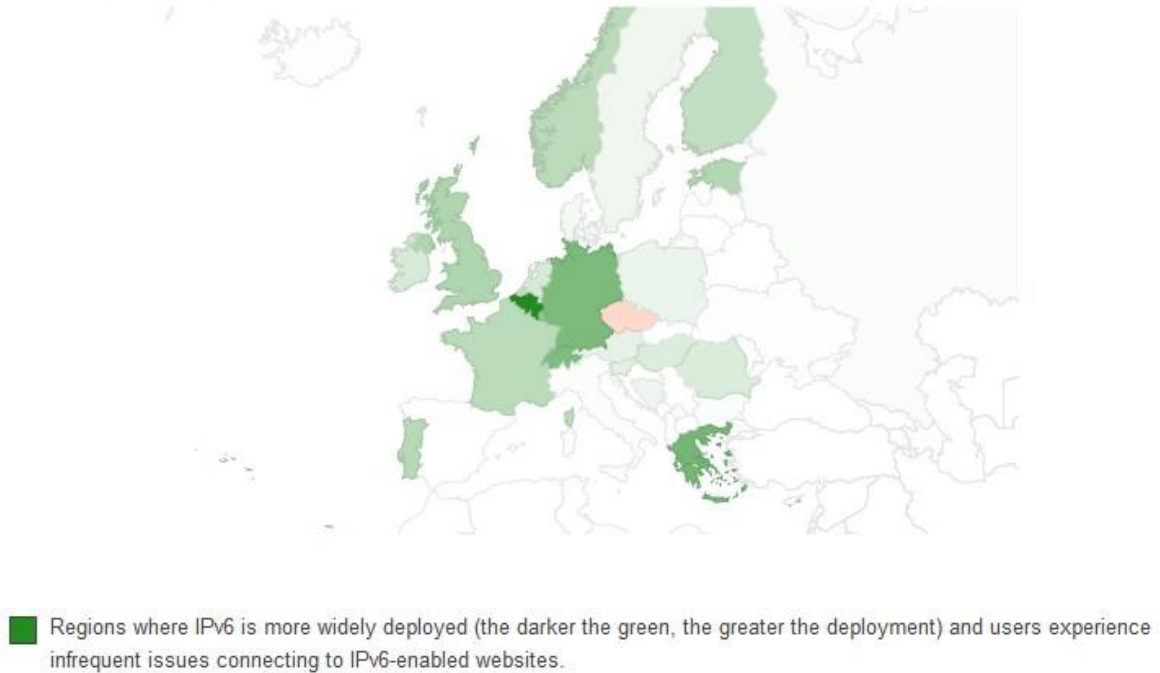


Figure 8: IPv6 deployment in Europe

Not every ISP provides you with IPv6 possibility, very few ISPs offer this feature. In this case, the network construction would look different to construction with IPv4. When asked how the

current situation is with IPv6, ISP's answer is that efforts are being expanded to the higher availability of IPv6. Sometimes you do not get an answer to the question from the ISP.

There is also the possibility to transport IPv6 traffic through the IPv4 network. 6to4 tunneling would be needed in combination with IPv6 tunnel broker. Tunnel broker provides you with network tunneling services. According to the error rate of the 6to4 implementation by using anycasts, various contents seemed to be badly accessible via IPv6. Therefore, the document RFC 7526 contains recommendations not to use 6to4 tunneling. [RFC7526]

For this reasons, the honeypot webpage will be only accessible via IPv4. This solution may be dealing with less unique addresses but IPv6 availability and reliability are currently not the strength of IPv6. In the future, when every clients will work with IPv6 address, identification will be easier.

The second important step for good webpage performance and for reachability of the webpage is meaningful Domain Name System or DNS. It makes a difference between `www.myinformationdusan.domain` and `www.dusanjaguzovic.domain`. DNS should be very meaningful and the content of the web page should be represented as well as possible through its DNS name. If the web page represents a company, the domain name should represent the company name, for example `YourCompanyName.com`. If the web page represents yourself, then `YourName.com` is a great option. If a real webpage with a real name exists then you can change the SLD or second level domain just adding a "-" character between first and last name, for example `www.dusan-jaguzovic.domain`. This is the reason why the second one will be used in this thesis. A lot of different domains exist, for example `.com` or `.at`. The domains `com` and `at` are not for free, there are predefined prices per month. Because the project should not involve any costs, free dynamic DNS domain called `ddnss.de` will be used.

Some dynamic DNS services will not be indexed by Google and therefore your web page would never be found by Google. Before you create dynamic DNS account, you should check if there are other sites from the same dynamic DNS provider indexed by Google. This also should improve the reachability of your web page.

3.1.4 Tempt visitors to visit

The first important aspect to tempt visitors to visit your webpage is already described in the chapter web page reachability. The DNS name of your honeypot web page should represent you.

Additionally, search engine result list ranking is very important. Academic research indicates that 91.5% of users visit only the first result page on Google. The first 3 results for Google search are results where most of activity happens. The top 3 results on Google capture 61% of all clicks. For these reasons, the ranking of your honeypot web page is very important. If somebody is looking for your name, your honeypot web page should be in the top 3 results. In this case, the probability for visiting the web page is much higher and you are able to get information about the visitor. The honeypot web page's www.dusanjaguzovic.ddnss.de ranking on Google is very good, at second place, figure 9 Google ranking.

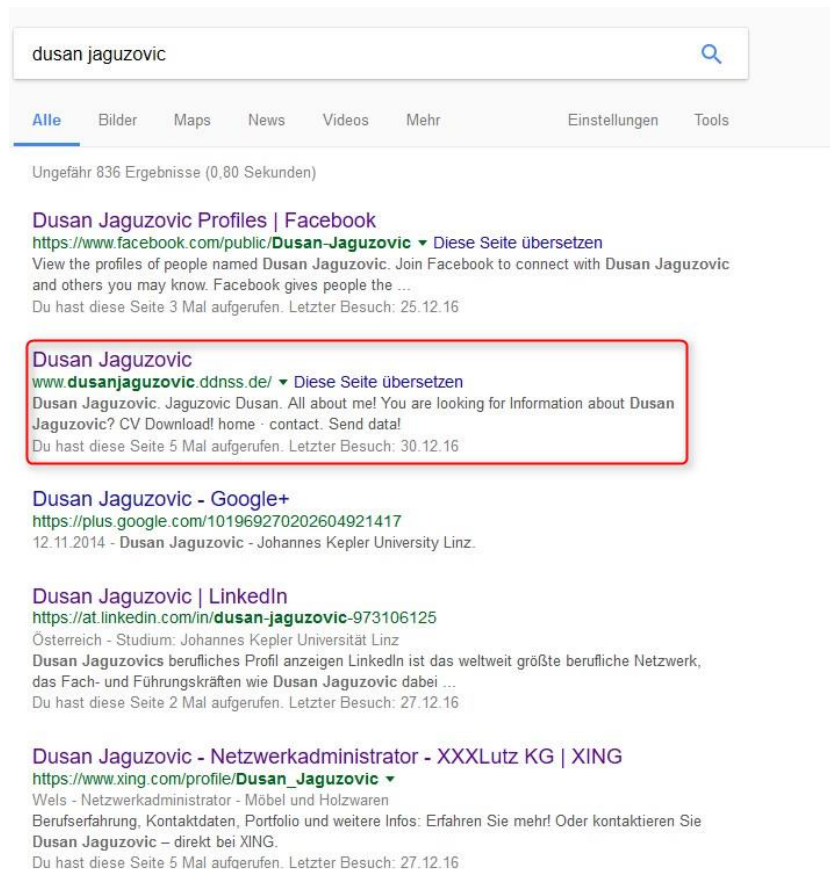


Figure 9: Honeypot google ranking

3.2 Social networks

Since their introduction, social network sites have attracted millions of users. Many of them use social networks in their everyday life. The number of social network users is increasing every day. In the year 2016 there are 2.34 billion social network users worldwide. Graphical representation of numbers of social network users is shown in the figure 10.

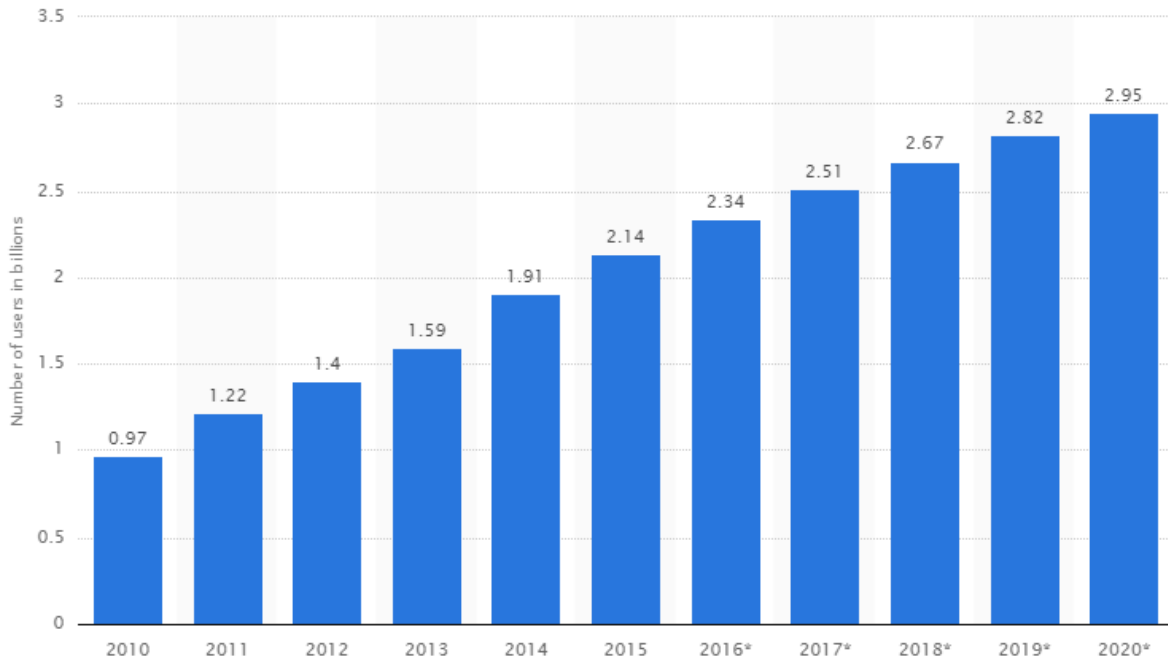


Figure 10: Social network users

The number of users is very interesting for this thesis. Social network profile could be an unambiguous identification of a person. Such a profile could be used to identify persons visiting your webpage. Such profiles are often providing you private information about user like address, age or photos. There two main focuses in this chapter. The first one is to find a way to find out if someone, visiting your webpage or looking for information about you, does have a social network profile. If yes, find out who is this person is.

How is it possible to get information that does not belong to you? How is it possible to do it legally and not to steal information? The usual approach to how a user visits and logs in on his social network profile is shown in the figure 11, the green path, directly to the application server. The goal is to attract the user to do the same thing but with an additional hop in the

route, the red path in the figure 11. The additional hop in the communication between client and application server will be the web server running in the environment where analysis will be done and where you have ability to access the data. In this case, it will be the honeypot server, described in the chapter before. Different APIs implemented on the honeypot server are able to do the scenario described in the figure 11 Login Social Networks. If the user uses the red path to login into application, you are able to evaluate the data. You get the application userID or the name of user and you are able to identify the person by its unique characteristics.

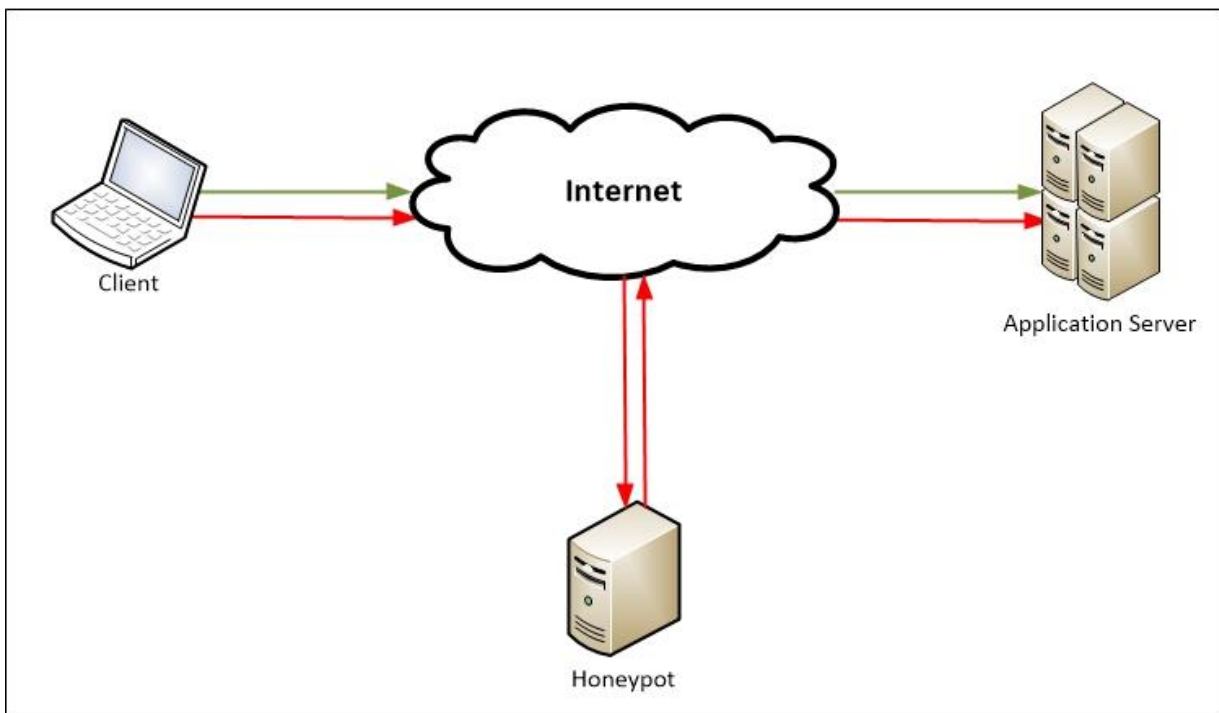


Figure 11: Social network login

The second main focus in this chapter is to find out if somebody visits your social network profile directly and not from external sources. This is a more complicated way because you do not have access to logging information on the application server.

3.2.1 Facebook

The focus in this chapter is on getting the users' Facebook information who are visiting your webpage or Facebook profile. Nowadays, Facebook is the most popular social network site worldwide. Facebook has 1.7 billions of active users [Stat 01]. At least every Facebook user is providing a name to the public. Additionally, a lot of them provide a profile photo, other photos and additional information about themselves.

As described before, there are two main focuses in this chapter. The first one is finding out which Facebook user is visiting your webpage using Facebook APIs. The second one is finding out who is visiting your Facebook profile directly.

The solution to the first question, to the first task is described in the chapter Facebook API. In this chapter we will deal with the second problem and we will investigate if there is possibility to solve this complicated problem.

The first limitation with Facebook is that there is no feature implemented and provided by Facebook to list all your visitors like Xing or LinkedIn. There is no possibility to be notified if someone just visits your profile. On the internet, you are able to find many applications for mobile clients and many computer software that claim to be able to find such information. In almost all cases these claims are false. Nobody, except authorities like FBI or CIA, has access to the data from Facebook and nobody is able to find this for you. Authorities get the data from Facebook.

In addition to this, a lot of Facebook users claim that there is a "hidden" option in the source code that provides you the list of people which visited your profile. It should be the list sorted by number of visits, its name is "InitialChatFriendsList". As the name already says, it is just a list of friends which are shown in the initial chat list on Facebook. This was also proven with a small test. The test scenario is simple, a new Facebook user without friends was created, in this case user A. Few other users were also created for testing purposes, in this test these are user B, user C and user D. These users visited the Facebook profile of the user A several times.

The “InitialChatFriendList” was still empty. Even if users B, C and D visit the Facebook profile of user A many times, the “InitialChatFriendList” remains empty.

The most promising possibility is the solution with recommended friends. Facebook and other social networks have to know who to recommend to another user. There are many criteria which are used to recommend somebody to somebody else. Few examples for criteria are:

- Mutual friends
- Place of residence
- School
- Common Facebook groups
- Employer
- Common events
- Actual location
- Common predictions
- Phone book
- Searches on Facebook [Inet 04]

Facebook’s algorithm in the background for calculating friend suggestions is always running. In some cases, one criteria is more important, the others less important. For example if you do not visit any events, the criteria “common events” is less important for your friend suggestions. The chance that you know friends of your friends are very good. The chance that you know people working for the same company are also very good. The chance that you know a person living in same town is also very good. Another thing which is important is your phone book if you are using a mobile device to be online on Facebook. Facebook wants to have access to lot of services running on a mobile phone. But the most interesting thing for this thesis are the searches on Facebook. It means, Facebook records everything a user does. If you search for a person, Facebook knows that you have a reason for looking for it. If Facebook uses all criteria to recommend you friends, the solution for the problem in this chapter is to limit Facebook just to use the last criteria, searches on Facebook.

If you have an existing Facebook profile with a lot of friends and lot of groups, it could be more complicated. But what will happen if you create new profile, with the same name, same picture without any friends, any groups and any other activities on this profile? Facebook does not know who to suggest to you because you do not search with this profile, you do not add friends, you do not “like” other photos, you do not publish your location and you do not publish your home town, your school or your employer. The next best choice will be the person who was looking for you on Facebook, based on Facebook searches recorded before. The conditions are that you do nothing with the new blank Facebook profile. Its name is the same as yours as well as other information about you that are interesting for other people except: place of residence or phone number. If you do something with this profile, Facebook will suggest you friends based on your activities. The only activity you should do is to check friend suggestions.

If somebody is looking for you on Facebook, he or she will probably search for your name. If you create a new blank profile with the same profile photo, searching person will not know which one is right and which one to visit. This person will probably visit both profiles, the real one and the new one for analysis. So if the new profile does nothing, the next best friend suggestion will be the person which was searching for you. The real profile will probably get many friend suggestions made based on other criteria described before.

Needed steps:

- Create new Facebook profile with same name
- Select the same profile picture for new profile as for original (Facebook does not make any restrictions if there are two identical profiles)
- No activities with new blank profile
 - No new friends
 - No new groups
 - No “likes”
 - No “shares”
 - No “comments”

- Check for friend suggestions regularly

Some tests were made and the results are presented in the test chapter.

There are a lot of limitations with Facebook. You as a user are very restricted. Everything you do on your Facebook page is restricted. You are not able to include your own JavaScript code like on Tumblr. You are not able to upload images with originating binary code, Facebook makes a picture of the picture you want to upload. There is also an option to upload PDF files. In this case, Facebook makes an image of every PDF page and uploads it as images, it means if you prepared a PDF file with tracking code in it, it will not work.

On the other hand, Facebook as an organization has a lot of possibilities to track you. Facebook stores a lot of data about you, not only what you say or who you talk to, but also what you like, where you have been and a lot of other personal information about you. According to tests made with a diagnostic tool called Abine DNT+, they noticed that Facebook has more than 200 trackers watching your behavior online. These trackers come in the shape of cookies, JavaScript, 1x1 tracking pixels, and Iframes. [SamF 01] Since Facebook is the most popular social network site worldwide, Facebook APIs are also very often used on external sites. For example the Facebook “Like” Button and Facebook Comment API are very often used on external sites. That’s also a way in which Facebook collects data about other people and how Facebook is able to track people even if they only visit external web sites. This is a great advantage of Facebook and other big companies like Google in terms of tracking and collecting private data. You can consider Facebook and Google as two great bridges connecting two islands, where millions of people go over the bridges every day. You can consider all other smaller companies and private websites as small wooden bridges and small boats, which use very few or none.

3.3 Career-oriented social networks

3.3.1 XING

Xing is the one of two most important career-oriented social networks but only in Europe. About 76% of all visitors are from Germany and about 90% of all visitors are from Germany, Austria or Switzerland. A career-oriented social network, as the name indicates, is a kind of social network whose primary purpose is career. The platform offers everybody the possibility to personalize their profile, to upload a profile photo or to be member of a group with a main focus on career.

There are two different memberships available [Xing 01]. The basic membership is free. The premium membership costs depending on the length of contract. Many core functions like searching people with specific qualifications or messaging everyone regardless of whether you are connected or not with them are only included in the premium membership.

There also exists a function called “Who viewed my profile” or “my visitors”. This function is only partially available for basic membership users. Premium members have more availabilities and possibilities to find out who viewed their profile. They can see the profiles of Xing users who visited their page. In addition to Xing user, premium members are also able to see who visited their profile from other websites, for example Google or other sites redirecting to their Xing profile, but also partially. They can see for example what the search term was to get to your Xing profile. There is very little documentation for premium membership. Additionally only 9% of all users are premium member. Premium member is not able to see that for example person John Doe visited his Xing profile redirected from google.com, except John Doe is Xing member and is logged in. Figure 12 shows how the visitors are represented in Xing.

If Xing member visited your profile, as basic member you are able to see the picture of this person regardless of whether he/she is a friend of you or not. Premium members retrieve more information about the visitor however premium membership is not free. 3 or 12 month subscription cost € 4.95/month or € 3.95/month. [Xing 02] Premium membership was not tested in this thesis.

In the figure 12 Xing you can see the second entry. There is one visitor who came from www.google.at. It means that this visitor came from www.google.at to your Xing profile but it does not automatically mean that that exact name was entered into google search. It is possible that other keywords that are visible to google for this Xing profile were entered. The keywords can be the employer or school education for example.

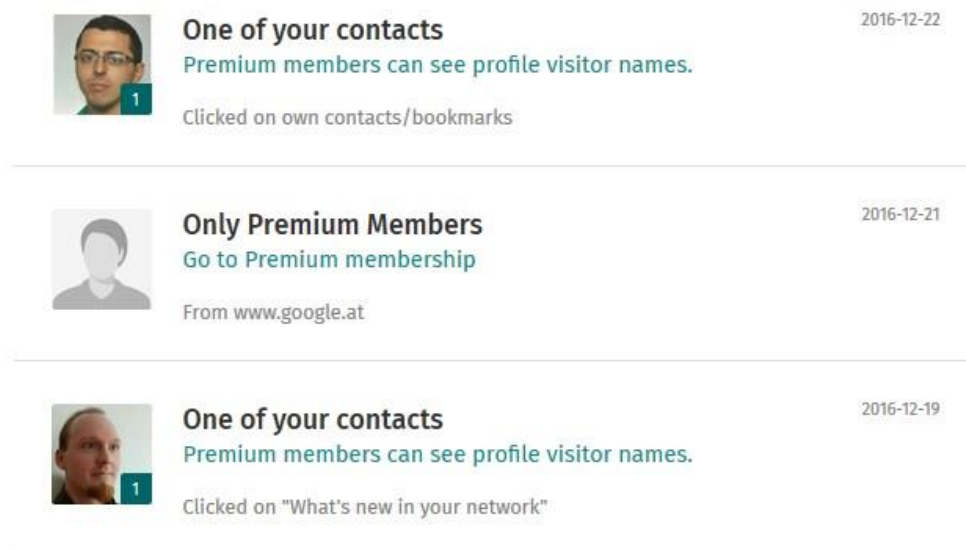


Figure 12: Xing visitors

The main question in this chapter is, is it possible to make a reverse search to find out who these visitors are. It is complicated for visitors coming from google or other referrers because of missing information and missing unique identifier. The only possibility is to combine Xing information with information obtained by other mechanisms described in this thesis. If one user visited your Xing profile on 21st of December 2016 and you have one log entry on the web server running the honeypot website with www.google.at as referrer, it is possible that the user visited your Xing profile and the user visited your honeypot website is the same. As shown in the figure 12, the limiting information is the missing time stamp on Xing. With a precise time, the comparison of Xing visit with log entry on the honeypot server would be much more efficient. The solution can be to check often, for example every 15 minutes.

In the case of Xing members visiting your Xing profile, there are two possibilities to find the user who visited your profile.

3.3.1.1 Identification based on user ID

This is the first and more precise method to identify the Xing user who visited your website without knowing anything about him/her except the information shown in the picture XY Xing.

Step 1: Right click on the profile image and copy the image address. The image address will look like this `https://www.xing.com/img/users/2/4/b/4c871bf38.12664672,1.64x64.jpg`. Paste the address of image into text editor.


Step 2: Identify the user ID in the image address. It is always the value behind the hexadecimal value, in the example in step 1 it is the value 12664672 behind the hexadecimal value 4c871bf38 separated by a dot.

Step 3: Combine the user ID found in step 2 with the following URL

`https://www.xing.com/events/widgets/organized/<user ID>` to get URL like this

`https://www.xing.com/events/widgets/organized/12664672` and paste the URL into the browser.

Step 4: The name associated with the user ID is displayed. Figure 13 shows the name of user who visited your Xing profile based on the URL from step 3.



Upcoming events organised by Milan Travar

(Currently no events available)

Figure 13: Xing example result

Step 5: As long as the name does not often occur on Xing, you can look for the name in the Xing search field and identify the appropriate profile.

This method is very accurate and delivers good results. More tests are in the test chapter.

3.3.1.2 Identification with google image search

The second and less precise method to identify Xing users, who visited your website, is with the help of google image search. Google image search is a feature provided by google for reverse image search. You can use a picture as your search to find related images from around the web. Google reverse image search results may include:

- Similar images
- Sites that include the image
- Other sizes of the image you searched for [Google 03]

Search using an image works best with pictures with the appropriate size and with good picture resolution. Searching with a small picture with size of 64x64 pixels will not work well. As you can see in the first method, identification based on user ID in step 1, at the end of URL there is a profile picture of the user who visited your site. On Xing it is always 64 pixels in the width and 64 pixels in the height. In addition to this, according to some studies, google image search is not able to do facial recognition. It only shows similar images that have similar combinations of shapes, proportions and colors to the original image. But the tests made with Xing images provide different results whose accuracy varies. Sometimes the result is very good but sometimes the results have no similarity with the original. [Inet 03].

The process of searching Xing user with help of image search feature looks as follows:

Step 1: Right click on the profile image and save the image.

Step 2: Open following URL <https://images.google.com/> and upload image from step 1.

Step 3: The search can be limited on certain website, in this case it should be limited to Xing. Add *site:xing.com* in the search filed of google image search after the image. This limitation increases the hit rate.

Some tests are made and results are presented in the test chapter.

3.3.2 LinkedIn

LinkedIn is a business and employment oriented social networking service. There are about 106 million active users on LinkedIn (September 2016) [Wiki 02]. This social networking service is mainly used for job searches and for job offers. According to Alexa Internet, a company that provides commercial web traffic data and analytics, LinkedIn is the 14th most popular website in the world in the year 2016.

To know who is visiting your LinkedIn profile is of great importance. For every LinkedIn user, LinkedIn provides an option called “Who’s Viewed Your Profile?”. With this option you are able to identify the person who visited your profile in the last 90 days by its LinkedIn name. There are two versions of your LinkedIn Profile. The first version is public version. In this case, every search engine is able to find you and every visitor is able to visit your profile. If this happens, you are not able to identify the visitor. The second version is the full one. In this case only registered LinkedIn user are able to find you and to visit your profile. Change the default version of your profile from public to full profile to limit the access to your profile.

You can use this option (December 2016) moving your cursor over “Profile” at the top of your homepage and select “Who’s Viewed Your Profile”. Another possibility is to click the number next to “X people viewed your profile”, next to your profile photo at the top of your homepage.

There is no other feature needed for LinkedIn. There are no limitations like time stamp or external users on LinkedIn. Everything you want to know about your visitors will be offered by the function “Who’s Viewed Your Profile”. Some tests will be made in the chapter test.

3.4 File Tracking

Monitoring and tracking files which contain information about certain a person could help to identify the searching person. Monitoring a file could tell you, for example, who opened or printed a document, when that was, and where somebody did it. There are some reasons for wanting to track files. One of them is maybe to check if somebody tried to open a document unauthorized. Another reason could be for example for statistical purposes, who opened a document most often and where this was, maybe to move the selling of a product to another region.

In this thesis, files from the honeypot server will be tracked and monitored with the aim of finding out who downloads the files. Files will contain fictitious data about a person. Fictitious data will be used because there is no need to distribute real information about you if you just want to find out who is looking for you.

Tracking the files does not mean that you can track the file just until next hop. For example, Server A provides B a PDF file for download. Person C downloads the B.PDF file and you are able to know person C's public IP address, person C's browser user agent or the time when person C downloaded the file. But what happens if person C sends this file to person D? We should be able to find out that person D received the file and opened it, too. Additionally, if person D sends the file to persons E, F, G and H, we should be able to know about this. But will we be able to detect the difference between sending and sending plus opening by persons E, F, G and H? Actually it is not as important to know that somebody received the file with my information as much as that they opened it. As you can see in the figure 14, it should be possible to find out whenever someone opens the document. The document "calls back home" and a log file entry on the honeypot server will be made. Then we will be able to know when and who opened the file. A logging example is also shown in the figure 14. The problem with this solution is that the user has to accept the access to external sources, in this case the URL of the honeypot where logs are made. There is one way to reduce this impact. Customization of the URL should make the pop up message less suspicious. Prepending of some random text, for example `www.google.com` or something similar could work. One possibility is to prepend

for example a username and password without meaning within a URL. One possibility is to prepend some text with google to your honeypot URL. One example is shown below. Username <google> and password <com> will probably make the pop up message less suspicious for most users.

```
http://google:com@dusanjaguzovic.ddnss.de/
```

Appending text to make the URL longer or to let the file viewer not show the message will work, and some tests are made and presented in the test chapter.

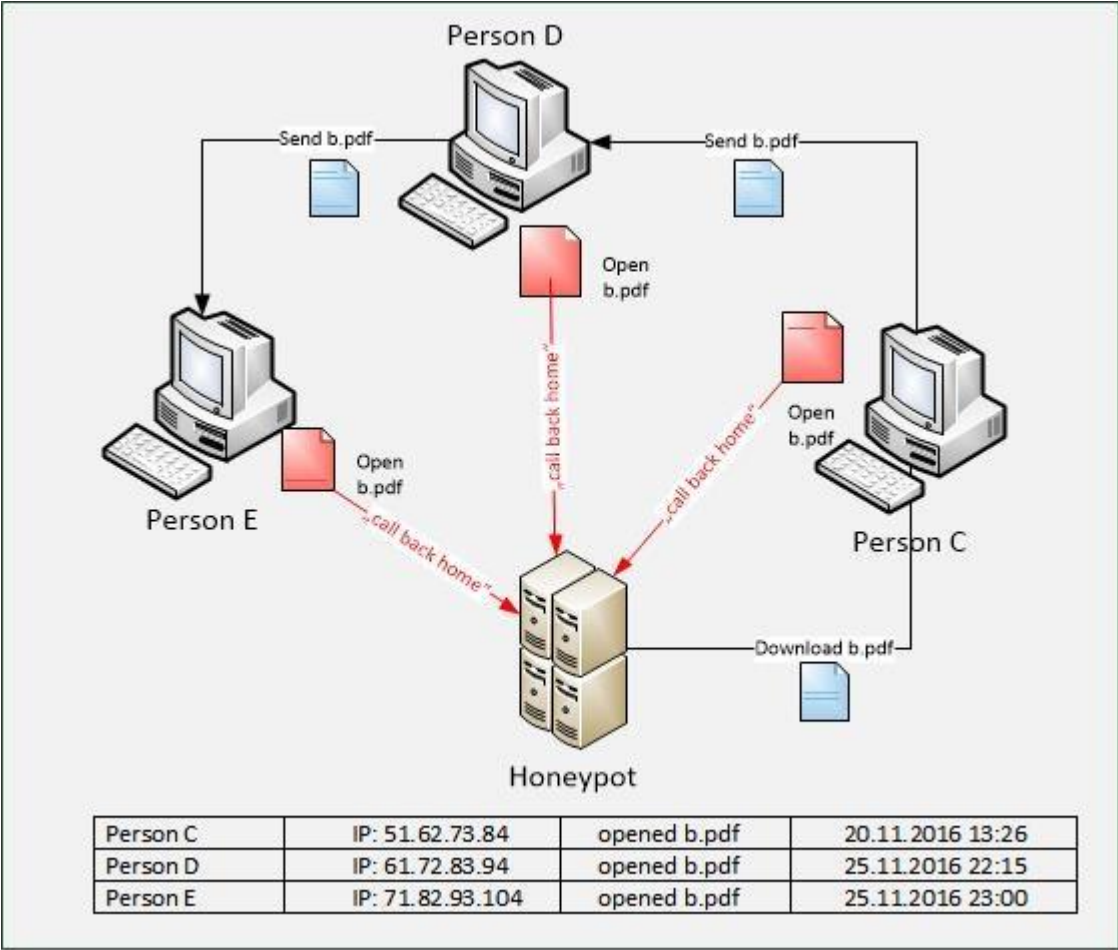


Figure 14: File tracking

3.4.1 PDF Tracking

Portable document format or PDF is a file format used to present documents independent of operating system or application software. PDF is the most used file format on the web, see figure 15. Every internet browser is able to show PDF files. There are a lot of real-time analytics and visit notification software which provide the ability to track PDF files. A disadvantage of these providers is that the prices grow up to \$ 5,000 depending on the type of licence. Some of the PDF tracking tools do not exist anymore, the rest are not suitable for the purpose of this thesis. They do not provide the useful information. Therefore, one's own tracking solution will be created in this thesis.

Scenario:

1. PDF files will be prepared for tracking
2. PDF files will be named [Person name]_Address.pdf and [Person name]_CV.pdf
3. Dummy PDF files will be available on the webpage for download
4. Monitoring and tracking mechanism will be able to make logs if a file is opened or printed

Point 1 is most the important step in the scenario. A PDF file is created. If somebody downloads the file, a log entry will be made automatically on the honeypot server. If someone opens the PDF file, you have to implement a kind of trigger to call back home. The simplest solution is to open a link directly from a PDF file, a log entry will be made automatically. If the opened link belongs to you and your honeypot web site, you are also able to see if someone opened the PDF file.

To implement such triggers in addition to a PDF viewer, you need software for editing PDF files. In this thesis, Adobe Acrobat Pro DC demo version will be used. The Demo version is valid up to 30 days free of charge. Adobe Acrobat Pro DC provides you also a feature called JavaScript for Acrobat, which allows to insert JavaScript in PDF files [Java 01].

In this case, you as the creator of PDF file, are able to create additional features in that PDF file. The solution for Step 1, step by step is as follows:

- Write the information you want to put into PDF file and save it as image
- Create a new PDF file with Adobe Acrobat Pro DC
- Place a button over the whole area. The probability that somebody clicks on the button in the PDF file is bigger than if you create a small button somewhere in the file
- Change the appearance of the button to the information you saved as an image in step one, image as background image
- Set JavaScript function to the button
 - o If somebody clicks the button in the CV PDF file call the page
<http://www.dusanjaguzovic.ddnss.de/googlesearchprocesssucess/googlesearchprocesssucess23/googlesearchprocesssucess123/hidden.html>
 - o If somebody clicks the button in the address PDF file call the page
<http://www.dusanjaguzovic.ddnss.de/googlesearchprocesssucess/googlesearchprocesssucess23/googlesearchprocesssucess123/addressDownloaded.html>
- A log entry will be made on the honeypot web site and you are able to know if somebody opens the file
- An immediate redirection is made from the hidden.html and addressDownloaded.html to www.google.com to make the process even less suspicious.

URLs are customized, they contain some random words that contain “google” to make the URL less suspicious.

The following source code is used to trigger the URL access:

```
var button1 = this.getField("Button1");  
button1.setAction("MouseUp",  
"app.launchURL('http://www.suport.google.at.dusanjaguzovic.ddnss.de/googles  
earchprocesssucess/googlesearchprocesssucess23/googlesearchprocesssucess123  
/hidden.html');");
```

The problem with this solution is that not every PDF viewer supports JavaScript. Table 4 shows which viewer supports JavaScript and which doesn't.

Viewer	JavaScript Support
Adobe Acrobat Reader	Yes
Adobe Acrobat Pro DC	Yes
NitroPDF	Yes
Firefox web browser	No
Microsoft edge web browser	No
Google chrome web browser	No
Internet explorer 11 web browser	Yes

Table 4: PDF Viewer JavaScript Support

If somebody opens the PDF file with a web browser, the probability for log entry on the honeypot web server is very low. The probability increases if somebody opens the PDF file with a PDF viewer like Adobe Acrobat Reader. To increase the probability that the user downloads and opens the file with a PDF viewer, download HTML attribute will be used. The download attribute specifies that the target will be downloaded when a user clicks on the hyperlink. You can see the implementation in the following HTML source code.

```
<a href="CV_Dusan_Jaguzovic.pdf" download="CD_Dusan_Jaguzovic.pdf">Dusan's CV </a>
```

All these measures are intended to bring the user to download the PDF file, to open the file with a PDF viewer that supports JavaScript and to accept the access to external URL. A log entry on the honeypot server is made when somebody opens a specific PDF file. This can be done with several files, too. The log entry example on the honeypot web page is shown below:

```
90.146.185.145 - - [08/Mar/2017:23:11:26 +0100] "GET /googlesearchprocesssucess/googlesearchprocesssucess23/googlesearchprocesssucess123/addressDownloaded.html HTTP/1.1" 200 "-" "Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:51.0) Gecko/20100101 Firefox/51.0"
```


3.4.2 DOC(X) Tracking

Another file format which is very common on the web is the DOCx file format, shown in the figure 15. For this reason you should also make this format available on your honeypot. Previously, there was only the DOC file format. This format is used by Microsoft Word versions up to 2007. The newer DOCx extension signifies the Office Open XML international standard for Office documents used by Word 2007 and later. The reason why the DOCx was developed is because Microsoft wanted to create a new file format which will be the standard. Other software products had issues reading DOC files. For this reason, DOCx was developed. DOCx encoding work was done in XML. On one hand, this is an advantage of DOCx file, but on the other hand, this is a disadvantage for tracking word files. Files with DOCx extension are not allowed to contain macros. The only possibility to call back home, as explained in the chapter 3.4, is to implement a macro with an auto open function. This function should access the external URL on the honeypot web server. A log entry will be made on the honeypot server and you are able to see when somebody opens the file. Because it is not possible to implement macros in a DOCx file, the DOCm or DOC file format will be used. DOCm file format is a DOCx document with macros.

Two DOC(m) documents will be prepared, one a "CV" and one an "Address" document. The visual basic code looks as follows:

```
Private Sub Document_Open()  
Set wshshell = CreateObject("WScript.Shell")  
wshshell.Run  
"www.dusanjaguzovic.ddnss.de/googlesearchprocesssucess/googlesearchprocesss  
ucess23/googlesearchprocesssucess123/addressDownloadedDoc.html"  
End Sub
```

The code above executes an access to the URL on the honeypot web server. There are two problems with this solution. The first problem is, the user has to activate macros. If the user does not activate macros, the function will not be executed. The second problem is, even if the user activates macros but the clients is offline or not connected to the internet, you will not get any log entry on the honeypot web server. There is no solution for these two problems.

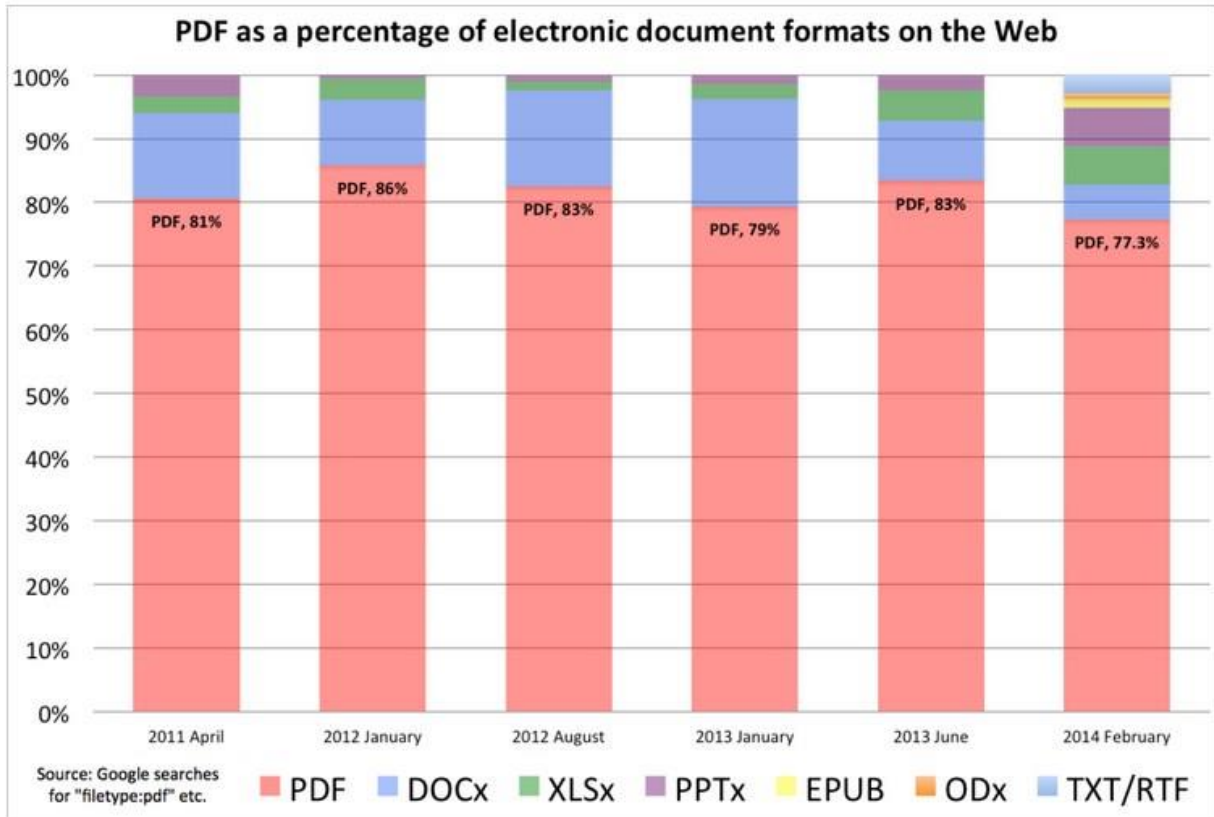


Figure 15: Most popular document formats on the web

In this case, the download attribute in the HTML source code specifies that the target will be downloaded when a user clicks on the hyperlink, the same code as for PDF file.

3.5 E-Mail

3.5.1 Received E-Mail

About 215 billion E-Mails are sent per day. [Stat 02] If you want to find out who is interested in your life, in your privacy, or in information about you, E-Mail analysis is also an important part. It is not only the analysis of received E-Mails that is important but also the right use of E-Mail addresses. Finding out who sent you an e-mail helps you to find out who is interested in you and contacted you directly for different purposes, for example if somebody sent you threatening e-mails.

Email tracking is a well-known method for tracking sent E-Mails. In this thesis, sent mails are not as interesting and important as received E-Mails. We want to find out who is sending E-Mails to us and not to whom we send a message.

Finding the source address of a received E-Mail can identify the sender. Every internet email message format is defined by RFC 5322 [RFC5322]. Every message consists of two major sections, the message header and the message body. The header is structured into fields such as "From", "To" or "Subject" and additional optional information about the message. The header field information will be also sent with the message and the receiver is able to see the header information. The body section contains the message. This section is less interesting for the investigation. The more important section is the header section.

Extracting the header section is different for different mail clients and for different mail services. It is different but it is possible for every E-Mail message to get the header information. The header block has exactly one header, which is structured into fields. Each field is specified by RFC 5322. There are many possible header fields but the most common fields related to message transport are: [Egli 01]

Header	Meaning
To:	E-mail address(es) of primary recipient(s)
Cc:	E-mail address(es) of secondary recipient(s)
Bcc:	E-mail address(es) for blind carbon copies
From:	Person or people who created the message
Sender:	E-mail address of the actual sender
Received:	Line added by each transfer agent along the route
Return-Path:	Can be used to identify a path back to the sender

Table 5: E-Mail header fields

Other header fields are:

Header	Meaning
Date:	The date and time the message was sent
Reply-To:	E-mail address to which replies should be sent
Message-Id:	Unique number for referencing this message later
In-Reply-To:	Message-Id of the message to which this is a reply
References:	Other relevant Message-Ids
Keywords:	User-chosen keywords
Subject:	Short summary of the message for the one-line display

Table 6: Additional E-Mail header fields

As you can see in the tables above, you are able to get a lot of information from the mail header which is available to everybody who receives an E-Mail. The most important header for the investigation and header analysis is the received field. It is a line added by each transfer agent along the route, it means you are able to recreate the route back to the sender. You read those from bottom to top as each server adds a new header to the list. The last hop in the route is the sender. Example of an email header is shown above.

```
Return-path: <user@example.com>
Received: from mac.com ([10.13.11.252])
  by ms031.mac.com (Sun Java System Messaging Server 6.2-8.04 (built Feb 28
  2007)) with ESMTTP id <0JMI007ZN7PETGC0@ms031.mac.com> for
  user@example.com; Thu,
  09 Aug 2007 04:24:50 -0700 (PDT)
Received: from mail.dsis.net (mail.dsis.net [70.183.59.5])
  by mac.com (Xserve/smtpin22/MantshX 4.0) with ESMTTP id 179BOnNS000101
  for <user@example.com>; Thu, 09 Aug 2007 04:24:49 -0700 (PDT)
Received: from [192.168.2.77] (70.183.59.6) by mail.dsis.net with ESMTTP
  (EIMS X 3.3.2) for <user@example.com>; Thu, 09 Aug 2007 04:24:49 -0700
Date: Thu, 09 Aug 2007 04:24:57 -0700
From: Frank Sender <sender@example.com>
Subject: Test
```

To: Joe User <user@example.com>
Message-id: <61086DBD-252B-46D2-A54C-263FE5E02B41@example.com>
MIME-version: 1.0 (Apple Message framework v752.2)
X-Mailer: Apple Mail (2.752.2)
Content-type: text/plain; charset=US-ASCII; format=flowed
Content-transfer-encoding: 7bit

Various online header analyzers are available online. In this thesis, Trace Email web tool will be used for analysis. [Tool 01] There are various IP address analyzers but [whatismyipaddress](#) analyzer delivers more information than the others. The added value of this application are the private IP addresses for example. An example of an output for the example email header looks something like what is shown in the figure 16. In the figure you can see the source IP address 70.183.59.6. Additionally, you see the result of geolocator which is also very helpful. Some tests are made in the test chapter.

The main limitations of this method are the mail clients. Depending on what mail client you use, the originating IP address will not be sent to your recipients. For example if somebody uses a web based mail client, in most cases the originating IP address can't be found. The first received field contains just the IP address of the used mail server. If you use a mail client like Outlook or Thunderbird, the originating IP address is very often set in the first received field. Detailed tests will be made in the test chapter. An additional limitation with this method is if the originating IP addresses are forged, typical use for spam or phishing emails. It is also possible that the spammer does not forge the IP addresses. In this case, the IP address will come from a machine in a botnet of which the actual owner is unaware. This should not be a problem because the most important part is not the spam and phishing emails but emails from real persons.

Additional improvement of identification based on emails is to use every email address only once. For every system, for every application, and for every web page a different email address should be used. It makes the organizational part more complex but it makes the reverse search much easier. If somebody writes you an email and uses [xy@hotmail.com](#) as destination address and you know you use this email address only for Xing, the probability that the user found your email address on Xing and visited your Xing profile is very high.

Analysis:

Return-path: <user@example.com>
Received: from mac.com ([10.13.11.252]) by ms031.mac.com (Sun Java System Messaging Server 6.2-8.04 (built Feb 28 2007)) with ESMTP id <0JMI007ZN7PETGC0@ms031.mac.com> for user@example.com; Thu, 09 Aug 2007 04:24:50 -0700 (PDT)
Received: from mail.dsis.net (mail.dsis.net [70.183.59.5]) by mac.com (Xserve/smtpin22/MantshX 4.0) with ESMTP id I79BOnNS000101 for <user@example.com>; Thu, 09 Aug 2007 04:24:49 -0700 (PDT)
Received: from [192.168.2.77] (70.183.59.6) by mail.dsis.net with ESMTP (EIMS X 3.3.2) for <user@example.com>; Thu, 09 Aug 2007 04:24:49 -0700
Date: Thu, 09 Aug 2007 04:24:57 -0700
From: Frank Sender <sender@example.com>
Subject: Test
To: Joe User <user@example.com>
Message-Id: <61086DBD-262B-46D2-A54C-263FE5E02B41@example.com>
MIME-version: 1.0 (Apple Message framework v752.2)
X-Mailer: Apple Mail (2.752.2)
Content-type: text/plain; charset=US-ASCII; format=flowed
Content-transfer-encoding: 7bit

Source:

The source host name is "www.whatismyipaddress.com" and the source IP address is 70.183.59.6.

Geo-Location Information

Country United States
State/Region CA
City Lake Forest
Latitude 33.6451
Longitude -117.6786
Area Code 949

Geo-Location Map



Figure 16: E-Mail header result

3.5.2 Sent E-Mail

In the chapter 3.5.1 received e-mails were investigated. In this chapter, we investigate how to track the sent e-mails. Is there any possibility to mark an e-mail and to be able track it, to be able to know when and where the e-mail was opened?

The most common way to do this is to send something in e-mail that calls back home and gives you the desired information. A few years ago it was possible to send a small 1x1 tracking pixel as image with every e-mail client. To do that you had to write an e-mail in HTML format and not in "text only format". Most companies send their newsletter written in HTML format. They are able to customize the e-mails. The next challenge is to embed a small 1x1 tracking pixel. This will be done in the body section of the HTML e-mail and will be done this way:

```

```

Every time or only first time user opens the received e-mail a log entry on the honeypot web server will be made. If every time or only first time depends on the mail client. If mail client provides the image caching feature, log entry will be made only one time. Every next time this image will be opened from the cache. If mail client does not provide the caching feature, log entry will be made every time user opens the mail because the image has to be loaded every time. Nowadays almost all mail clients prevent sending 1x1 tracking pixels. Microsoft and Google make changes to the HTML code of the e-mail. If you embed the image and send it, the receiver receives the image but with completely different source code. The image will be loaded only during the creation of the mail, can be seen in the log files on the honeypot web server. This is exactly what we want to avoid in this chapter.

One possible solution is to send the e-mail without any mail client. Build a SMTPS connection, port 465, to the mail server and send SMTP e-mail with the embedded tracking pixel. To do this we use swiftmailer, a PHP library for sending e-mails from PHP applications.

```

<html>
<head>
</head>
<body>
<?php
//require_once 'swift/lib/swift_required.php';
require_once 'C:/swiftmailer-swiftmailer-81fdccf/lib/swift_required.php';
$transport = Swift_SmtpTransport::newInstance('smtp.gmail.com', 465, "ssl")
    ->setUsername('dulevw@gmail.com')
    ->setPassword('password');

$mailer = Swift_Mailer::newInstance($transport);

$message = Swift_Message::newInstance('Test Subject')
    ->setFrom(array('dulevw@gmail.com' => 'ABC'))
    ->setTo(array('dulevw@hotmail.com'))
    ->setBody('<img src=http://dusanjaguzovic.ddnss.de/images/Tracking.gif>',
'text/html');

$result = $mailer->send($message);
?>
</body>
</html>

```

In the source code shown above, you can see the PHP code for sending an e-mail with embedded tracking pixel in the body section. The connection to the mail server, smtp.gmail.com, is established first. Username and password of existing e-mail account is needed, too. In this case an e-mail will be sent from dulevw@gmail.com to dulevw@hotmail.com. The same code will be used for testing purposes in the test chapter. The tracking pixel is hidden in the file Tracking.gif and is located on the honeypot webserver.

3.6 Tumblr

Tumblr is a blogging platform where you can share texts, images and videos. There are about 550 million Tumblr users monthly, about 280 million blogs and about 53 million new posts daily. These big numbers are the reason why we investigate Tumblr.

Is there any possibility to track visitors of your Tumblr profile? Yes, there is a very efficient way to do this. Using StatCounter you are able to see a lot of information about visitors. StatCounter is a web traffic analysis tool. It is a simple but powerful real-time web analytics service that can track, analyze and understand your visitors. StatCounter works on a lot of platforms, figure 16.



Figure 17: StatCounter platforms

The first step is creating a StatCounter account. Then you are able to combine StatCounter with all of the platforms shown in the figure 17.

If you want to combine StatCounter to track your Tumblr visitors, StatCounter creates a JavaScript tracking code. The code, shown below, has to be inserted into Tumblr, “Website theme’s” description.

```
<!-- Start of StatCounter Code for Tumblr -->
<script type="text/javascript">
var sc_project=11213584;
var sc_invisible=0;
var sc_security="5d7d875c";
var scJsHost = (("https:" == document.location.protocol) ?
"http://secure." : "http://www.");
document.write("<sc"+"ript type='text/javascript' src='" +
scJsHost+
"statcounter.com/counter/counter.js'></"+"script>");
</script>
<noscript><div class="statcounter"><a title="free hit
counter" href="http://statcounter.com/" target="_blank"></a></div></noscript>
<!-- End of StatCounter Code for Tumblr -->
```

StatCounter uses client side JavaScript to grab information about your visitors and their browsers, operating systems, screen resolutions and so on. This information is then sent back to the back-end architecture where it is stored and analyzed. Visitor information you get from StatCounter is very meaningful. You can see an example result in the figure 18. There were two visits on January 8th. Both visitors are from Linz in Upper Austria and have Liwest as ISP. You can also see the public IP address, operating system, user agent and screen resolution. Additionally you can see the referring links. This information is very helpful for identification.

Date	Time	Referring Link
8 Jan	20:02:15	(No referring link)
8 Jan	19:55:37	https://www.tumblr.com/dashboard
8 Jan	19:56:11	https://dulevw.tumblr.com/
8 Jan	19:56:13	https://www.tumblr.com/dashboard
8 Jan	19:56:13	www.statcounter.com/ (Exit Link)
8 Jan	19:56:25	https://dulevw.tumblr.com/
8 Jan	19:56:30	https://www.tumblr.com/dashboard
8 Jan	19:56:30	https://dulevw.tumblr.com/

Figure 18: Tumblr example result

StatCounter only works with platforms shown in figure 16. The question has to be asked, why it works on Tumblr and not on Facebook? Unlike many other simplified website development services, Tumblr allows you to include custom script tags, it means you are able to include your JavaScript code into Tumblr. You can include your own code or you can include external source code. It may be a security hole but Tumblr allows you to do this. This is exactly the reason why StatCounter does not work on Facebook Xing for example. They do not allow you to include your own code. The permissions are very limited for security reasons.

4 Tests

4.1 Honeypot Test

In this chapter practical experiments with Honeypot web pages are described. All other tests made in this chapter cover only one area, for example chapter 4.2 covers only Facebook investigations. Honeypot test covers more than one area. Honeypot web page log entries are combined with other results to get more accurate results at the end and to be able to identify someone easier.

As described in the chapter 3.1.3 and 3.1.4, search engine ranking is very important. At the time of tests, if somebody searched for the Name “Dusan Jaguzovic” or “Jaguzovic Dusan”, the Honeypot webpage was in 2nd place on the first page of google search results.

20 persons from different professional and life environments participated in the Honeypot test. They were just told to search for online information about a person called “Dusan Jaguzovic”. They should write down every step they do and associated time stamp giving me the availability to compare the logs made on the Honeypot web server and to identify them. At the end every participant should deliver the results he found, where and when he found this information. This should give me availability to get more precise results, too. The test group consisted for example of mechanics, computer scientists, medical staff, and an insurance broker. The test scenario should be as real as possible, for these reasons people with different occupations were chosen. In the table 7 you can see test results.

Total number of participants	20
Users used Google as search engine	20
Users visited Honeypot web site	20
Users visited only first Google result page	16
Users downloaded PDF file	6
Users downloaded DOC file	2
Users opened PDF with PDF Viewer which supports JavaScript	3
Users opened DOC file with Microsoft Word	0
Users activated GPS location	2
Users logged in with Google	2
Users logged in with Facebook	3
Users visited Xing profile with same referrer	20
Multiple visits	4

Table 7: Honeypot results

4.1.1 Interpretation of results

All participants which looked for the name “Dusan Jaguzovic” or “Jaguzovic Dusan” visited the Honeypot web page. The result might have been different if the Honeypot web page would be on the second or third Google result page. Where you can’t make a distinction is whether the visitor was looking for “Dusan Jaguzovic” or “Jaguzovic Dusan”. In this case, this is less important because in both cases the ranking of the Honeypot web page is same.

The very interesting thing is that all of the users used Google as the search engine, there was nobody who used Yahoo! or Bing. The only difference is that somebody used www.google.at, somebody used www.google.ba. The fact that all users found the Honeypot web page and visited it independent of country and search language is very important. Searching with www.google.at and searching with www.google.de do not always produce the same results [Inet 04]. Searching with Safari on an apple device and searching on Windows device do not always produce the same results. Few participants used apple devices, they also found the

Honeypot web page. 80% of all participants did not visit the second result page. It means, you as Honeypot web page provider have to keep the web page up to date and to strive for the best Google ranking. A short test was made, where the Honeypot web page was offline for a few days. The Google ranking had greatly deteriorated during this offline time period. After a longer time period, Google removed the Honeypot web page because it was not able to index the page. After a few days online, the web page was available on Google again.

A very interesting number is also the number of Xing visitors with the same referrer. There are a lot of matches. All 20 participants visited the Honeypot web page and the Xing profile from the same external source address within a few minutes. For example, there is a log entry on the Honeypot web page at 18:00 o'clock with referrer www.google.at. For the same data, one external visitor visited the Xing profile with the same referrer. In this case, the probability that it was the same user is very high. According to participants, the result evaluation was 100% correct. A great disadvantage of Xing in this case is that you get only the date of the visit and not the time. To reduce the error rate, you should evaluate the Xing logs in shorter time intervals. Another interesting thing is that not a single participant has a Xing account or was logged in when he visited my Xing profile.

A third of all users downloaded one of two PDF files and about 10% downloaded the DOC file. Only two persons of six which downloaded the PDF file, opened it with a PDF viewer that supports JavaScript. In this case, two log entries were made on the web server. You can see example log entry for viewing the CV file in PDF format.

```
178.115.129.190 - - [06/Mar/2017:21:14:11 +0100] "GET  
/googlesearchprocesssucess/googlesearchprocesssucess23/googlesearchprocesssucess123/hi  
dden.html HTTP/1.1" 200 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:51.0)  
Gecko/20100101 Firefox/51.0"
```

A very small percentage of participants logged in with their Google or Facebook account. One possibility is that they did not have an account. Another possibility is that they did not want to log in because they did not see any advantages or they thought it unimportant.

4.2 Facebook Test

In this chapter some experiments with Facebook are realized. The challenge is to find out if there is a connection between users on Facebook who are not friends. In addition to this it is also about extracting useful information. The main task of Facebook tests is to verify if there is useful information that some unknown Facebook profiles visited your own profile. This is realized with two unconnected Facebook profiles. The most important thing is that you do not need any additional Facebook features, you do not have to implement code on your own website or you do not have to have programming skills.

Agents	User A	User B	User C	User D
Profile name	Dusan Jaguzovic	Jaguz Dusanovic	Tobias Dreher	Patrick Clarkson
Profile URL	https://www.facebook.com/dusandule.jaguzovic	https://www.facebook.com/people/Jaguz-Dusanovic/100014174527675	https://de-de.facebook.com/people/Tobias-Dreher/100014736010318	https://de-de.facebook.com/people/Patrick-Clarkson/100014652584895
Facebook user since	2008	13. November 2016	26. December 2016	26. December 2016
Friends	865	0	0	0

Table 8: Facebook test users

User B's, C's and D's profiles were created for testing purposes. These users are searching for user A. User A's goal is to find hints for it and extract useful information, like their profile names.

The worst case for Facebook test is when user B, C or D visits user A's profile and user A is not able to find any indications or any traces for this behavior.

The best case for Facebook test is when one of test users visits user A's profile and user A is directly or indirectly able to find it and identify them. Facebook notifications are a direct way that somebody was on your Facebook profile. An indirect way is to come to a conclusion through different information shown by Facebook, for example friend or group suggestions.

Test Case 1	No Friend Suggestion
Agent under Test	User A, User B
Description	User A has a lot of Facebook friends, user B has no friends. User A does not know that user B exists. There are no relationships between user A and user B in form of friendship or common groups. User A is an active user and is online daily. User B is online only few times a month.
Scenario	User A does not visit user B's Facebook profile. User B does not visit user A's Facebook profile. Both of them know nothing about each other. Both do their usual activities on Facebook.
Purpose	Verify that user A does not get friend suggestions for user B from Facebook.
Expected result	Facebook does not suggest user B as friend to user A.
Result	Facebook does not suggest user B as friend to user A.

Table 9: Facebook Test Case No Friend Suggestion

Test Case 2	Friend Suggestion
Agent under Test	User A, User B
Description	User A has a lot of Facebook friends, user B has no friends. User A does not know that user B exists but user B knows that user A exists. There are no relationships between user A and user B in form of friendship or common groups. User A is an active user and is online daily. User B is online only a few times a month.
Scenario	User A does not visit user B's Facebook profile. User B visits user A's Facebook profile at regular intervals, 5 times.
Purpose	Verify that user A gets friend suggestions for user B from Facebook based on user B's visits on user A's profile.
Expected result	Facebook suggests user B for friend to user A.
Result	Facebook suggests user B for friend to user A.

Table 10: Facebook Test Case Friend Suggestions

In the Figure 19 you can see the friend suggestion which means user B was suggested for a friend to user A just after 5 times visiting the profile of user A. There are no mutual friends.



Figure 19: Facebook Dusan's friend suggestions

Additional test result is shown in the figure 20, Patrick's Suggestions, Jaguz Dusanovic is suggested for friend to Patrick Clarkson, a new blank Facebook profile. The Facebook algorithm calculated the suggestion based on Jaguz Dusanovic's 5 visits on Patrick Clarkson's profile.

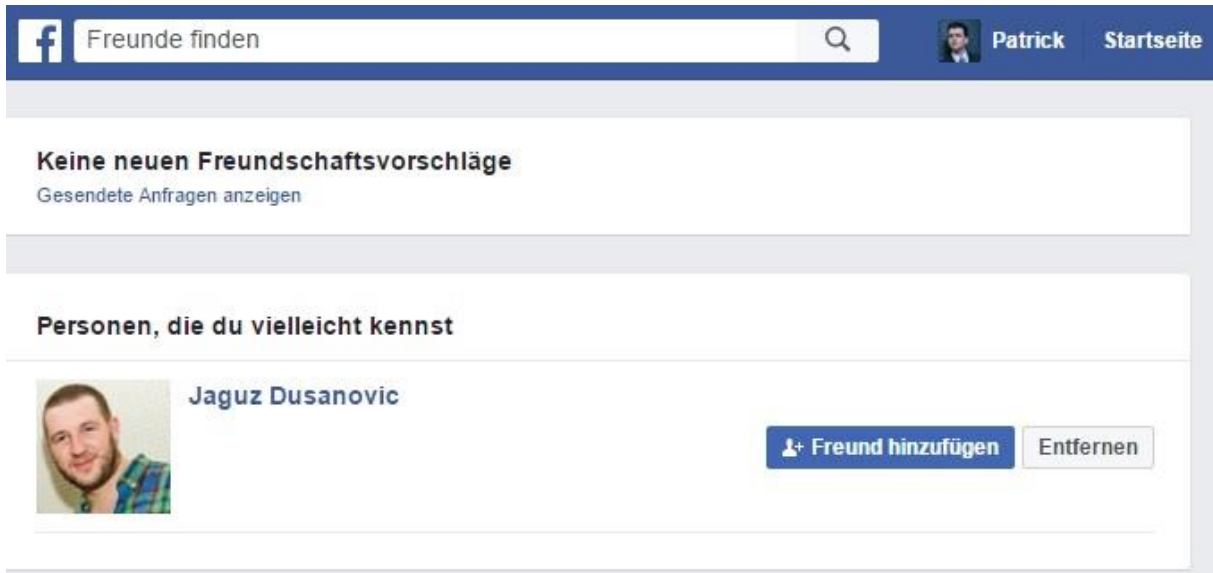


Figure 20: Facebook Patrick's friend suggestion

This solution for finding out who is visiting your Facebook profile works fine. You can see additional test results in the figure 20 to 21. Test case 2 made with few different Facebook users with completely different names. Only disadvantage with this solution is that you do not always immediately get friend suggestions. It may take a few days until Facebook makes you a suggestion. Additional preventive measure you should do is to disable the option that unregistered users are able to visit your Facebook profile. With this measure you are not able to find out who wants to know something about you but you limit that only registered users can see your profile. In this case you have a solution to find out which registered user visited your profile.



Figure 21: Facebook Tobias's friend suggestions

In the figure 21, you can also see “Laura Schmidt” as friend suggestion. In this case, you can be sure that this person visited the Facebook profile of “Tobias Dreher” because there are no events visited by “Tobias Dreher”, there are no mutual friends and there are no common interests because Tobias Dreher is a completely new profile without any activities except activities in the previous tests.

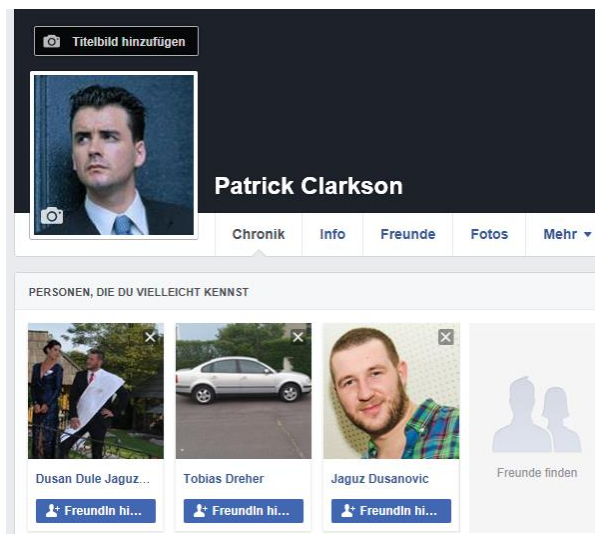


Figure 22: Facebook Patrick's friend suggestions

4.3 XING Test

In this chapter there are two methods that will be tested. The first one is the better one and more accurate method with the user ID. The second one is less accurate, but will be tested too.

For testing purpose the same scenario will be used for testing both methods, see chapter 3.3.1. An attempt will be made to find the same person in two different ways. For testing purposes the existing Xing profile of Dusan Jaguzovic, in this test user A, will be used for investigation. An additional new profile for testing is also needed. The new profile's name is Patrick Clarkson and will be used to test both methods if Xing visitor and visited user are not connected.

Test Case 1	User ID Test
Agent under Test	User A, User B
Description	User A is Xing member. User B is also Xing member and visits the profile of user A.
Scenario	User A is a basic member of Xing and does not have features of a premium member. User B visits Xing profile of user A directly from Xing platform. User A has to find out who the visitor is, using the first method, user ID method from chapter Xing. Visitor B can but must not be connected with user A on Xing.
Purpose	Find the name of person who visited the profile of user A.
Expected result	Result accuracy should be 100%. Every Xing visitor should be identified.
Result	According to Xing and to tests made, every visit on the profile is being logged. Every Xing visitor is identified by his Xing profile name. Accuracy is 100%.

Table 11: Xing Test Case User ID Test

Visitor #1:



One of your contacts 2016-12-22
Premium members can see profile visitor names.
Clicked on messages

Figure 23: Xing visitor entry #1

The result is:


Upcoming events organised by Milan Travar

(Currently no events available)

Figure 24: Xing visitor result #1

The result for visitor #1 is Milan Travar, true.

Visitor #2:



A XING member Today
Premium members can see profile visitor names.
Searched for your name

Figure 25: Xing visitor entry #2

The result is:

Upcoming events organised by Patrick Clarkson

(Currently no events available)

Figure 26: Xing visitor result #2

The result for visitor #4 is Patrick Clarkson, true. Patrick Clarkson is not one of the contacts of user A, you can recognize this by the label “A XING member” and not “One of your contacts”.

The accuracy of this method is 100% as expected, but only in the case that the visitor has no default profile picture. The limitation of this method is when the Xing user does not have a profile picture or has a default profile picture, figure 27.

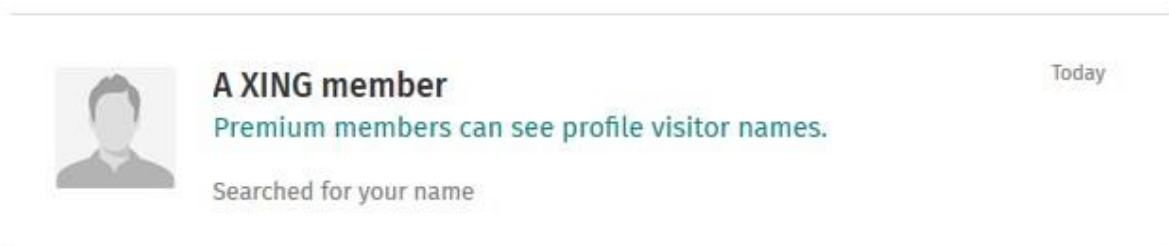


Figure 27: Xing no profile picture visitor

The result you get when you copy the link of the profile picture looks as follows

https://www.xing.com/assets/frontend_minified/img/users/nobody_m.64x64.jpg. As you can see there is no user ID and you are not able to find out who the visitor is.

Test case 2 does not differ much from the test case 1. Only difference is the identification method. In test case 2, identification is performed through an image search. Desired accuracy in this test case is only half of the first desired accuracy. The reason is explained in the chapter 2.3.

Test Case 2	Image Search Feature
Agent under Test	User A
Description	User A is a basic member of Xing. No additional features are activated.
Scenario	User A is a basic member of Xing and is not a premium member. Visitors visit Xing profile of user A directly from Xing platform. User A has to find out who the visitors are using the second method, google image search method from chapter Xing. Visitors must not but can be connected with the user A.
Purpose	Find the name of person who visited the profile of user A.
Expected result	Result accuracy should be more than 50%. As many users as possible should be identified.
Result	Results from this test case made with 4 different persons reach 50% accuracy. It can vary and depends on several factors, like image quality or reverse image search engine.

Table 12: Xing Test Case Image Search

Visitor #1:



Figure 28: Xing visitor entry #1, method 2

The result is:

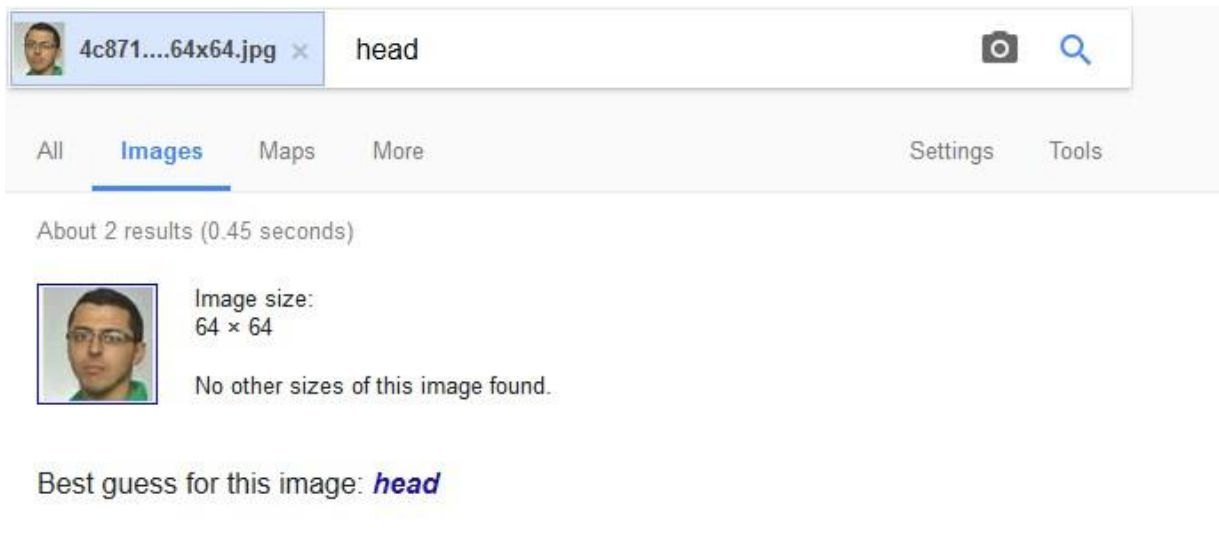


Figure 29: Xing visitor result #1, method 2

The result for this image search is **head**. Google image search was not able to identify this person by its real name or its real Xing profile.

Visitor #2:



Figure 30: Xing visitor entry #2, method 2

The result is:

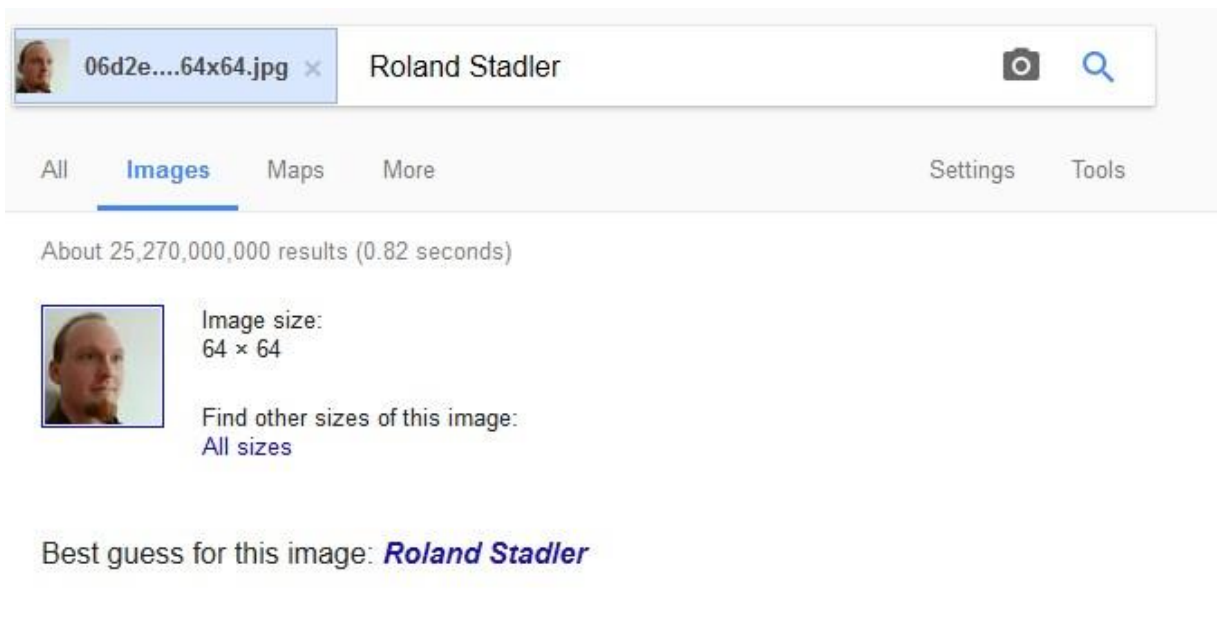


Figure 31: Xing visitor result 2, method 2

The result for this image search is Roland Stadler. In this case this is really Roland Stadler and Google image search was able to identify this person by its real name.

Visitor #3:

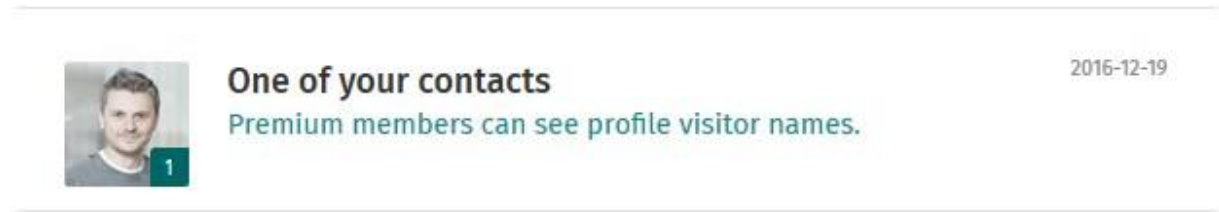


Figure 32: Xing visitor entry #3, method 2

The result is:

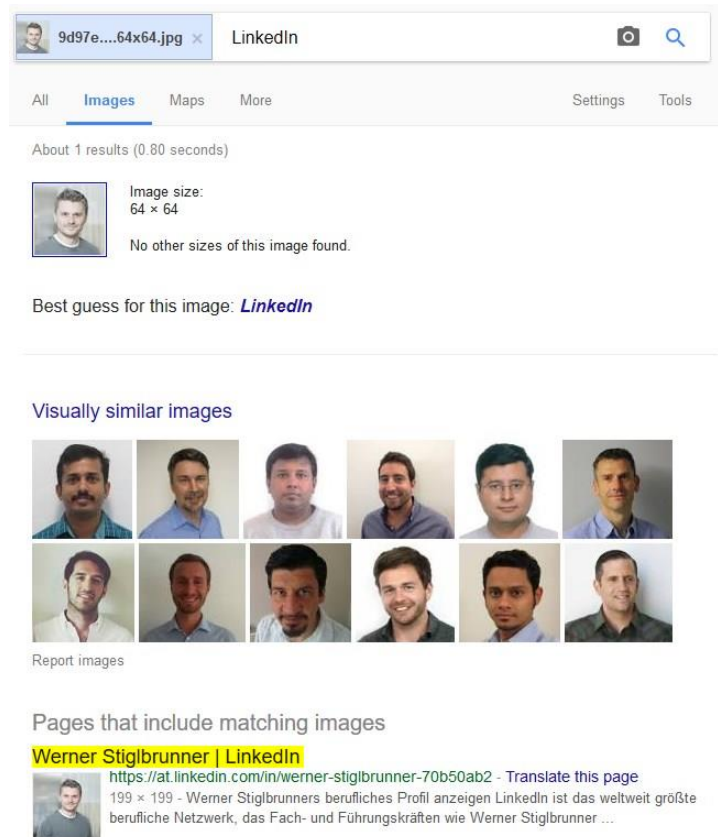


Figure 33: Xing visitor result #3, method 2

The result for this image search is Werner Stiglbanner. In this case this is really Werner Stiglbanner and the google image search was able to identify this person by its real name even if you limit the search to LinkedIn and not to Xing. The profile picture is the same.

Visitor #4:



Figure 34: Xing visitor entry #4, method 2

The result is:

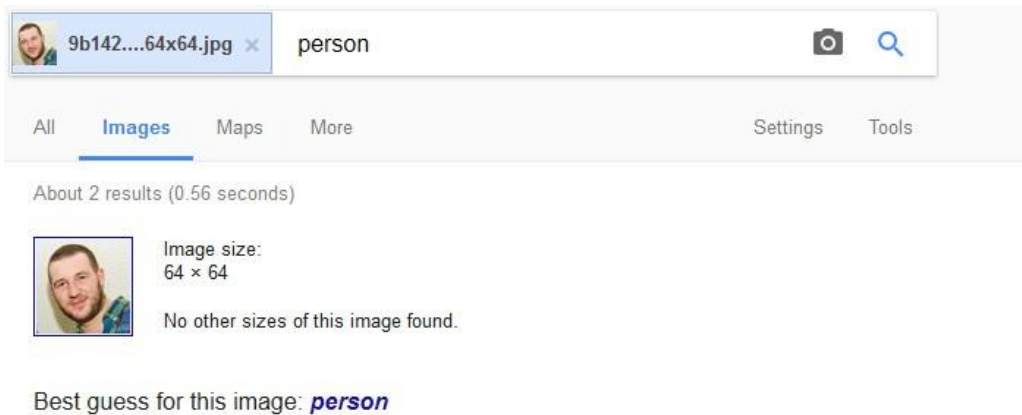


Figure 35: Xing visitor result #4, method 2

The result for this image search is **person**. Google image search was not able to identify this person by its real name or its real Xing profile.

The accuracy of this method is 50%, 4 searches divided by 2 hits. The limitation for this method is also when the Xing user does not have a profile picture or has a default profile picture. Additional limitation is the google image search feature with its accuracy of about 50%. Even with a bigger picture and a picture with a better resolution, google image search was not able to identify visitor #4. There are few other reverse image search engines but all of them had worse results, for example Bing Image Match or Yandex. Tineye was tested with a hit rate of 0. [Tool 02]

As explained at the beginning of the chapter, it is more complicated to find out who the visitors are who come from external sources, for example from www.google.at. In this test, we will try to find out as much information as possible about visitors coming from external sources. There is no direct way to do this. Combination of information obtained by other mechanisms in this thesis can be helpful.

One possible scenario could be that the external Visitor is searching for your name on google. First he visits your Xing profile, Xing records this visit. You are able to see that somebody visited your Xing profile coming from www.google.at on 28th of December. Then the same user goes back to google search results and visits your homepage, your honeypot website. A log entry will be created also on the honeypot website, that user with a certain public IP address visited your website on 28th of December with www.google.at as referrer. If no other log entries exist on Xing and on honeypot log for the same date with the same referrer, the evidence suggest that this is likely to be the same person. Checking the log files at shorter intervals can improve the results, this means even if you have more than one visit per day, it is possible to identify the person if you get the information at the right time.

Test Case 3	External visitors
Agent under Test	User A
Description	User A is basic member of Xing. No additional features are activated.
Scenario	Visitors visit Xing profile of user A from external sources. User A has to find out who the visitors are using the information from Xing and combine them with information obtained by other mechanisms for example honeypot logs.
Purpose	Find as many information as possible about the external visitor.
Expected result	Find a person who is most likely the right one.
Result	With help of honeypot, almost all external visitors could be identified.

Table 13: Xing Test Case External Visitors

External visitor #1:



Only Premium Members
Go to Premium membership

From www.google.at

Today

Xing entry from
December the 28th, 2016

Figure 36: Xing external visitor

Log entry from honeypot:

```
62.101.148.138 - - [28/Dec/2016:18:41:50 +0100] "GET / HTTP/1.1" 200
"https://www.google.at/" "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/55.0.2883.87 Safari/537.36"
```

The result:

There is no other entry on the web server for the same date with referrer www.google.at. In this case, we have indices that the Xing visitor and honeypot website visitor could be the same person.

4.4 LinkedIn Test

As explained in the chapter LinkedIn, there is only one option which covers all visitors of your LinkedIn profile, “Who’s Viewed Your Profile”.

At first, proof that nobody is able to visit and view your LinkedIn profile if he or she is not logged in. There are two versions of your LinkedIn Profile. The first version is the public version. In this case, every search engine is able to find you and every visitor is able to visit your profile. If this happens, you are not able to identify the visitor. The second version is the full one. In this case only registered LinkedIn user are able to find you and to visit your profile. If you want to view LinkedIn profile of somebody else, and you are not logged in and the profile of this person is not public, you get the message to log in.

The second step is the proof that you can identify all visitors of your LinkedIn profile. If nobody can visit your profile from external sources and you are able to see only visitors from LinkedIn, it means you are able to identify every visitor. As shown in the Figure 37 LinkedIn Visitors, Patrick Clarkson visited the profile on the day the function was used. You can see all visitors for last 90 days. There are no additional limitations with LinkedIn.



Figure 37: Who's viewed your LinkedIn profile

4.5 E-Mail Test

4.5.1 Received E-Mail

Analyzing email headers using the web tool “Whatsmyipaddress” is very helpful. The advantage of this tool over other tools is that you get additional information like user’s private IP addresses. Private IP addresses can be helpful to hint the difference between two identical public IP addresses. For example, if two different persons send you an email from their company work stations, the public IP address will be the same, but the private IP address will change. These addresses are marked as ignored by the tool because they are private, but they could be also very important for the identification.

Analysis 1:

Description: Payroll email sent from a company with more than 20.000 employees to one his employee.

Mail from: lohnzettel@lutz.at

Mail to: dulevw@hotmail.com

Result:

Analysis:

```
Received: from DB3PR02CA0028.eurprd02.prod.outlook.com (10.242.134.38) by V11PR0202MB2766.eurprd02.prod.o
Received: from inbound.mail.protection.outlook.com (213.199.180.150) by DB3PR02CA0028.outlook.office365.com (10
Received: from AM5EUR02FT005.eop-EUR02.prod.protection.outlook.com (10.152.8.59) by AM5EUR02HT189.eop-EI
Authentication-Results: spf=none (sender IP is 91.90.150.48) smtp.mailfrom=lutz.at; hotmail.com; dkim=none (message
Received-SPF: None (protection.outlook.com: lutz.at does not designate permitted sender hosts)
Received: from SNT004-MC5F14.hotmail.com (10.152.8.51) by AM5EUR02FT005.mail.protection.outlook.com (10.152
X-IncomingTopHeaderMarker: OriginalChecksum::UpperCasedChecksum::SizeAsReceived:1504;Count:15
Received: from s2048.ppsmtp.net ([91.90.150.48]) by SNT004-MC5F14.hotmail.com over TLS secured channel with Mi
Received: from pps.filterd (s2048.ppsmtp.net [127.0.0.1]) by s2048.ppsmtp.net (8.16.0.17/8.16.0.17) with SMTP id uAT
Received: from srv00-ex2010-01.lutz.gmbh ([212.152.255.71]) by s2048.ppsmtp.net with ESMTP id 26xy6e8e38-2 (vers
Received: from exchangebe1.lutz.gmbh (100.0.150.150) by SRV00-EX2010-01.lutz.gmbh (10.10.21.58) with Microsoft :
Received: from lohnsoan3 ([10.10.101.8]) by exchangebe1.lutz.gmbh with Microsoft SMTPSVC(6.0.3790.4675); Tue, 28
IP address 10.10.101.8 was ignored because it is a Private-Use Network address.
```

Figure 38: Mail Test 1 Result

As shown in the figure 23, you can see that the source IP address is 212.152.255.71, first blue IP address from bottom. This is the public IP address of sender. Additional important IP address is 100.0.150.150, the green one. This is the private IP address of the exchange server of the XXXLutz company, its public IP address is 212.152.255.71. Another private IP address is 10.10.101.8, the orange one, which belongs to lohnsan3 client, which is probably the client from whom the mail was sent. Even if the public IP address is shown as source IP, you are able to see the real private IP addresses which can help you identifying people sending you emails.

Geolocator is also able to locate the IP address. Geolocator tool CQCounter [Tool 03] is used and delivers information which is correct, based on personal knowledge. It belongs to the XXXLutz Zentrale, Roemerstrasse 39, 4600 Wels, AT. This mail was really sent from this location and all information and IP addresses are correct. The network infrastructure is known to the author of this thesis from his work and this mail was investigated for testing purposes, see figure 39.

212.152.255.71 - Whois Information

```
% This is the RIPE Database query service.
% The objects are in RPSL format.
%
% The RIPE Database is subject to Terms and Conditions.
% See http://www.ripe.net/db/support/db-terms-conditions.pdf

% Note: this output has been filtered.
%       To receive output for a database update, use the "-B" flag.

% Information related to '212.152.255.32 - 212.152.255.63'

% Abuse contact for '212.152.255.32 - 212.152.255.63' is 'hostmaster@at.tele2.com'

inetnum:        212.152.255.32 - 212.152.255.63
netname:        UTA-200805271
descr:          XXXLutz Zentrale, Roemerstr 39, 4600 Wels
country:        AT
tech-c:         UIO1-RIPE
admin-c:        UIO1-RIPE
status:         ASSIGNED PA
mnt-by:         AS8437-MNT
created:        2009-10-05T10:52:33Z
last-modified:  2009-10-05T10:52:33Z
source:         RIPE # Filtered
```

Figure 39: XXXLutz whois result

Analysis 2:

Description: Three loan requests were sent to Bank Austria. Three replies from Bank Austria were obtained. Two requests were sent to Bank Austria branch in Wels and one was sent to Bank Austria headquarter in Vienna.

Mail from: mail1@unicreditgroup.at; (Wels)

mail2@unicreditgroup.at; (Wels)

mail3@unicreditgroup.at; (Vienna)

Mail to: dulevw@hotmail.com

Results:

Mail from mail1@unicreditgroup.at

Analysis:

```
Received: from DB6PR0201CA0001.eurprd02.prod.outlook.com (10.168.49.139) by V11PF
Received: from inbound.mail.protection.outlook.com (213.199.180.145) by DB6PR0201CA
Received: from AM5EUR02FT035.eop-EUR02.prod.protection.outlook.com (10.152.8.58)
Authentication-Results: spf=pass (sender IP is 162.25.24.134) smtp.mailfrom=unicreditgrou
Received-SPF: Pass (protection.outlook.com: domain of unicreditgroup.at designates 162.2
Received: from BAY004-MC3F23.hotmail.com (10.152.8.58) by AM5EUR02FT035.mail.pro
X-IncomingTopHeaderMarker: OriginalChecksum::UpperCasedChecksum::SizeAsReceived
Received: from mx03.bagis.at ([162.25.24.134]) by BAY004-MC3F23.hotmail.com over TL
Received: from BAMC4PWMR2.mc04.unicreditgroup.eu (unknown [10.63.66.88]) (using T
IP address 10.63.66.88 was ignored because it is a Private-Use Network address.
Received: from BAMC4PWX6.mc04.unicreditgroup.eu ([10.63.66.43]) by BAMC4PWMR
IP address 10.63.66.43 was ignored because it is a Private-Use Network address.
From: RIEDER Thomas <Thomas.Rieder@unicreditgroup.at>
To: "dulevw@hotmail.com" <dulevw@hotmail.com>
Subject:
Thread-Index: AdI7WpdBSUlogh2OT82ghU33FJAH5w==
Date: Thu, 10 Nov 2016 13:59:05 +0000
Message-ID: <C5D62293755D5444AB65902F0BC967B41FF04583@BAMC4PWX6.mc
Accept-Language: en-US
Content-Language: de-DE
X-MS-Has-Attach: yes
X-MS-TNEF-Correlator:
x-originating-ip: [10.58.230.8]
IP address 10.58.230.8 was ignored because it is a Private-Use Network address.
```

Figure 40: Mail Test 2 Result 1

Mail from mail2@unicreditgroup.at

Analysis:

```
Received: from VI1PR02CA0008.eurprd02.prod.outlook.com (10.162.7.148) by VI1PR0202ME
Received: from inbound.mail.protection.outlook.com (213.199.154.111) by VI1PR02CA0008.ou
Received: from AM5EUR03FT020.eop-EUR03.prod.protection.outlook.com (10.152.16.80) by .
Authentication-Results: spf=pass (sender IP is 162.25.24.134) smtp.mailfrom=unicreditgroup.at
Received-SPF: Pass (protection.outlook.com: domain of unicreditgroup.at designates 162.25.24.134 as permitted sender)
Received: from COL004-MC3F21.hotmail.com (10.152.16.55) by AM5EUR03FT020.mail.protection.outlook.com
X-IncomingTopHeaderMarker: OriginalChecksum::UpperCasedChecksum::SizeAsReceived:148
Received: from mx03.bagis.at ([162.25.24.134]) by COL004-MC3F21.hotmail.com over TLS session
Received: from BAMC4PWMR9.mc04.unicreditgroup.eu (unknown [10.63.66.121]) (using TLSv1.2, cipher TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256)
IP address 10.63.66.121 was ignored because it is a Private-Use Network address.
Received: from BAMC4PWX6.mc04.unicreditgroup.eu ([10.63.66.43]) by BAMC4PWMR9.mc04.unicreditgroup.eu
IP address 10.63.66.43 was ignored because it is a Private-Use Network address.
From: JUNGMAIER Othmar <Othmar.Jungmaier@unicreditgroup.at>
To: "dulevw@hotmail.com" <dulevw@hotmail.com>
Subject: Per E-Mail senden: Objektcheckliste_LIEBE_L_ETW
Thread-Topic: Per E-Mail senden: Objektcheckliste_LIEBE_L_ETW
Thread-Index: AdI8BZKYhJHYHhgCRZe9gm0KGvRIFg==
Date: Fri, 11 Nov 2016 10:22:58 +0000
Message-ID: <C12A25DAB58CB44F96737EEB59538BE71FF06621@BAMC4PWX6.mc04.unicreditgroup.eu>
Accept-Language: de-DE, en-US
Content-Language: de-DE
X-MS-Has-Attach: yes
X-MS-TNEF-Correlator:
x-originating-ip: [10.58.230.5]
IP address 10.58.230.5 was ignored because it is a Private-Use Network address.
```

Figure 41: Mail Test 2 Result 2

The first two emails were sent from employee in the branch in Wels, Upper Austria. According to email header analysis, the source IP address is 162.25.24.134 which belongs to UniCredit Business Integrated Solutions Austria GmbH. There are also few private IP addresses which were ignored. The last private IP address for each is 10.63.66.X. If we look to the email header of email from Vienna, the last private IP address is 10.100.50.43. Two branches from Wels have the same last hop before going onto the Internet and headquarters from Vienna has another private IP address.

Mail from mail3@unicreditgroup.at

Analysis:

```
Received: from DB5PR0201CA0007.eurprd02.prod.outlook.com (10.164.231.145) by VI1P
Received: from inbound.mail.protection.outlook.com (213.199.154.112) by DB5PR0201CA0
Received: from AM5EUR03FT058.eop-EUR03.prod.protection.outlook.com (10.152.18.80)
Authentication-Results: spf=pass (sender IP is 162.25.24.134) smtp.mailfrom=unicreditgroup
Received-SPF: Pass (protection.outlook.com: domain of unicreditgroup.at designates 162.2
Received: from COL004-MC1F38.hotmail.com (10.152.18.54) by AM5EUR03FT058.mail.pr
X-IncomingTopHeaderMarker: OriginalChecksum::UpperCasedChecksum::SizeAsReceived:
Received: from mx03.bagis.at ([162.25.24.134]) by COL004-MC1F38.hotmail.com over TLS
Received: from BAMC4PWMMR12.mc04.unicreditgroup.eu (unknown [10.100.50.43]) (using
IP address 10.100.50.43 was ignored because it is a Private-Use Network address.
Received: from BAMC4PWX5.mc04.unicreditgroup.eu ([10.63.66.91]) by BAMC4PWMMR1
IP address 10.63.66.91 was ignored because it is a Private-Use Network address.
From: LUKIC Danijela <Danijela.Lukic@unicreditgroup.at>
To: "dulevw@hotmail.com" <dulevw@hotmail.com>
Subject: Ihr Finanzierungswunsch
Thread-Topic: Ihr Finanzierungswunsch
Thread-Index: AdJBsQckJs0U3LneTWtd9oq/pP0w==
Date: Fri, 18 Nov 2016 15:33:11 +0000
Message-ID: <E106017AD9050C41A3F1A9DCE3B5401C18EB992B@BAMC4PWX5.m
Accept-Language: en-US
Content-Language: de-DE
X-MS-Has-Attach: yes
X-MS-TNEF-Correlator:
x-originating-ip: [10.58.230.7]
IP address 10.58.230.7 was ignored because it is a Private-Use Network address.
```

Figure 42: Mail Test 2 Result 3

There is other useful information provided by email header. In the figure 43 below you can see that the sender used an android mail app to send the mail.

```
Content-Type: multipart/mixed; boundary="--_com android_email_121496934779160"
```

Figure 43: Used E-Mail app

Additionally, you are sometimes able to know the user's mobile provider. According <http://cqcounter.com/whois/> this IP address belongs to Hutchison Drei Austria GmbH.

```
X-Originating-IP: [178.115.250.211]
```

Figure 44: Public IP Address belongs to certain provider

4.5.2 Sent E-Mail

In addition to analyzing received emails, tracking sent emails has a lot of advantages. As described in the chapter 3.5.2 some tests are made with tracking sent emails. For testing purpose, Hotmail account with default settings was used. Using the default settings is very important, because sometimes there are additional features that should avoid loading tracking pixel. The purpose of the test is to see how email with default settings is being loaded.

For testing purposes, an email without text is sent. A tracking pixel is embedded in the body section of the email as shown in the chapter 3.5.2.

Log entries when an email is opened with a web based mail client two times:

```
90.146.185.145 - - [23/Apr/2017:15:22:32 +0200] "GET /images/Tracking.gif HTTP/1.1" 200  
"https://outlook.live.com/" "Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:52.0)  
Gecko/20100101 Firefox/52.0"  
  
90.146.185.145 - - [23/Apr/2017:15:23:31 +0200] "GET /images/Tracking.gif HTTP/1.1" 200  
"https://outlook.live.com/" "Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:52.0)  
Gecko/20100101 Firefox/52.0"
```

If this email is opened with an iPhone, it will create only one log entry because of the iPhone's default mail settings. The default setting for loading remote images is enabled and therefore the image will be loaded. If you open the same mail a second time, no entry will be created because the image will be loaded from the cache. In this case it depends on iPhone.

```
90.146.185.145 - - [23/Apr/2017:15:37:46 +0200] "GET /images/Tracking.gif HTTP/1.1" 200 "-  
" "Mozilla/5.0 (iPhone; CPU iPhone OS 10_3_1 like Mac OS X) AppleWebKit/603.1.30 (KHTML,  
like Gecko) Mobile/14E304"
```

4.6 Test summary

Test	Result
Honeypot test	See table 7
Facebook test case 1: No friend suggestion	No friend suggestions by Facebook if no visits on the profile
Facebook test case 2: Friend suggestion	Facebook suggests the persons for friend which visited your Facebook profile
Xing test case 1: User ID test	Identification of visitors which are Xing members and were logged in at the time of visit with accuracy of 100%
Xing test case 2: Image search feature	Identification of persons with the help of reverse image search with accuracy of 50%. The accuracy refers only to the tests made in this case.
Xing test case 3: External visitors	Identification of all external visitors with the help of honeypot and Xing visitors feature was possible.
LinkedIn test case 1	Find the LinkedIn person which visited your LinkedIn profile was possible and the accuracy is 100% (if LinkedIn shows you all visitors).
Received email analysis 1	Investigating received email from a known company revealed a lot of useful information about sender (geolocation, public IP range, address)
Received email analysis 2	Find out what the benefits are when you know the sender's private IP address. The result is that in test case you can make a difference between user's geographical locations.
Sent email analysis 1	Open sent email tracking pixel with web based email client and with default settings. The tracking pixel was opened and downloaded every time a user opens an email.
Sent email analysis 2	Open sent email tracking pixel with a mobile device and with preinstalled app on it. The tracking pixel was opened only once. The second time it was opened from the cache.

Table 14: Test summary

5 Summary

The internet is continuously growing. The amount of private information about individuals on the internet is growing proportional to the internet. Controlling the access to private data is becoming increasingly difficult. There are many reasons for these difficulties. Many tools have been developed to anonymize the user's behavior on the internet. A lot of other limitations exist that make the online identification more difficult and more complex.

It is very important to have control over private data and to know who is looking for you and who is interested in what information about you. On the one hand there are many risks if somebody wants to know everything about an individual. On the other hand there are many advantages of knowing that somebody searched for you on Google or visited your social network profile. Some benefits are based on job and some benefits are maybe based on your love life. Knowing that somebody visited your Xing profile indicates that somebody is maybe interested in your career.

The main question in this thesis was to find out if it is possible to find out who is looking for you online. If this is possible, how accurate are the solutions and what are their limitations?

One of the main focuses of research and implementation of this thesis were search engines. Millions of searches per day on Google and other search engines were reason enough to investigate the identification of people who are looking for me with the help of search engines. The main part for this investigation is the honeypot web page. A honeypot web page should serve as bait for all who are looking for me. Therefore the search engine result ranking is very important. As can be seen from the results in the chapter 4.1, the most important ranking is Google ranking. All test participants used Google as search engine. Another very important thing about honeypot web page are the implemented features like browser fingerprinting or geolocating. The features should help you identify those who are looking for you. Individual features may produce less accurate results but the combination of all features should produce the best result. Not only the combination of implemented features is an advantage but also the combination of the honeypot web page with third party services is also important. As can

be seen in the chapter 4.2, the combination of Xing and the honeypot web page is of great importance for identifying external visitors. The results from the chapter 4.1 show that you are able to find out if somebody is looking for you with the help of a search engine. Maybe in some cases you are not able to find out who this someone is because of for example anonymization tools but you have indications that somebody was looking for you.

Another very important fields of investigation for this thesis were social networks. On the one hand there are normal social networks and on the other hand there are career-oriented social networks. The biggest and best known social network today is Facebook with billions of members. Facebook is a part of investigation in the chapter 3.2. There are a lot of ways and a lot of tools described online and provided online for download that claim to be able to find out who is visiting your Facebook profile. Several of them were investigated but without success, for example "InitialChatFriendsList". The only way to do this is described in the chapter 3.2.1. Some tests were made in the chapter 4.2 and they confirm the assumptions from the chapter 3.2.1. The only one who knows all visitors on a Facebook profile is Facebook himself. Getting this information legally from Facebook is impossible. There are few limitations with the presented solution in the chapter 3.2.1 but it works very well and very accurately. The presented solution in the chapter 3.2.1 that works with Facebook can also be used for other services where the basic principle of friendship proposal is the same. Identification is not only possible with help of same friendship proposal but also with profile images and other data which can be used for identification.

The most famous career-oriented social networks were investigated, too. The ways to do this are presented in the chapter 3.3. All investigated and tested ways are very accurate and work very well. There are pre-implemented features that can be used for the purposes of this thesis. Some settings should be changed to limit the access only to those you can identify. Adjusting the settings improves the chance to identify the visitor of your career-oriented social network profile.

Another also important fields of investigation were files and e-mails. According to the results from the chapter 4.1, there are only very few visitors who download and open different files

from the internet. You can use file tracking because on the one hand you can't lose anything and on the other hand if somebody opens the file with the right software, you are able to get information needed for identification. Even if the number of those who have done this is very low you are able to identify these small percentage of people.

In the last chapter 4.5, both sent and received mails were investigated. Finding out who really sent you an email could be also very important if somebody is interested in you. More interesting is the tracking of sent emails with the help of tracking pixel. You can response to received emails to verify extracted data or you can track a new email you want to send. A lot of email clients downloads the tracking pixel from the honeypot webserver using the default settings. Several of them do this only once but the others do this every time, for example Hotmail web based email client. In this case you are able to get information about user which sends or receives the email you investigated or tracked.

5.1 Conclusion

Finding out who is looking for you online is very difficult and it becomes more and more difficult with new security and anonymization inventions. Many factors complicate the identification like tor network. Many anonymizations tools exist today that make the identification more difficult. Every day more and more people are using the internet and more and more people want to know what happens with their private data, who is able to see them and who actually saw them. To achieve that despite many limitations and many restrictions from the side of third party services, despite many anonymization tools and despite many other limitations, you should limit the access only to data that is under your control or partially under your control. You should limit the access to services where you are able to find out who was looking for you. There are many ways to do this described in this thesis. Additional you should put baits online so that you can be found easily. If somebody takes the bait, you are able to find a lot of information about the visiting client that helps you to identify the client. If you want to know who is visiting your profile, services like Instagram where you are in no way able to find out who is looking for you and who is visiting your profile, it should be avoided. Being careful with sharing information online is an important thing, too. Even if you are able

to identify the machine that is looking for you or visiting your profile, it is very hard if not impossible to identify the person working with the machine.

5.2 Outlook to future

As described in the chapter 3.1.3, using IPv6 in the future would have many advantages. Identification results would be more accurate because of IPv6's more uniqueness. Google could also facilitate the identification of users. By announcing the database or parts of database of geolocation positions of a lot of wireless routers, described in the chapter 3.1.2.1.2, the identification results would be more accurate. The best help for identification online would be new features for different services, like the "Who viewed my profile" feature of Xing. There will be for sure more anonymization tools in the future. There will be a lot of different ways how to make IP spoofing with IPv6, but still it will be easier to identify people who are looking for you.

6 Bibliography

- [Unesco 01] Toby Mendel, Andrew Puddephatt, Ben Wagner, Dixie Hawtin, Natalia Torres
Global Survey on Internet Privacy and Freedom of Expression
United Nations Educational, Scientific and Cultural Organizations, 2012
- [Inet 01] Internet live stats, last request: April 04, 2017
<http://www.internetlvestats.com/google-search-statistics/>
- [Wiki 01] Various authors, Wikipedia, free encyclopedia, last request: April 04, 2017
https://en.wikipedia.org/wiki/Google_Alerts
- [PewRes 01] Lee Rainie, Sara Kiesler, Ruogu Kang, Mary Madden,
Anonymity, Privacy, and Security Online
September 5, 2013
- [Inet 02] Tina Sieber, January 31, 2010, last request: April 04, 2017
<http://www.makeuseof.com/tag/find-googling/>
- [Google 01] Google Analytics, last request: April 04, 2017
<https://www.google.com/analytics/>
- [Spitzner 01] Lance Spitzner
Honeypots: Tracking Hackers
Addison Wesley, September 13, 2002
- [EFF 01] Peter Eckersley
How Unique Is Your Web Browser?
Electronic Frontier Foundation

- [Google 02] Google Statistics, last request: April 04, 2017
<https://www.google.com/intl/en/ipv6/statistics.html#tab=per-country-ipv6-adoption&tab=per-country-ipv6-adoption>
- [RFC7526] O. Troan, B. Carpenter
Deprecating the Anycast Prefix for 6to4 Relay Routers
Internet Engineering Task Force (IETF), May 2015
- [Stat 01] The statistics portal, last request: April 04, 2017
<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>
- [SamF 01] Samantha Felix, September 9, 2012, last request: April 04, 2017
<http://www.businessinsider.com/this-is-how-facebook-is-tracking-your-internet-activity-2012-9?IR=T>
- [Xing 01] Xing, last request: April 04, 2017
<https://faq.xing.com/en/my-projobs-my-premium-upgrade/what-memberships-are-available>
- [Xing 02] Xing, last request: April 04, 2017
https://www.xing.com/upsell/premium_offers?from_hub=true&utm_campaign=buffer&utm_content=buffere3908&utm_medium=social&utm_source=plus.google.com
- [Google 03] Google Support, last request: April 04, 2017
<https://support.google.com/websearch/answer/1325808?hl=en>
- [Inet 03] Mat McGee, June 17, 2011, last request: April 04, 2017
<http://searchengineland.com/up-close-with-google-search-by-image-82313>

- [Wiki 02] Various authors, Wikipedia, free encyclopedia, last request: April 04, 2017
<https://en.wikipedia.org/wiki/LinkedIn>
- [Stat 02] The statistics portal, last request: April 04, 2017
<https://de.statista.com/statistik/daten/studie/252278/umfrage/prognose-zur-zahl-der-taeglich-versendeter-e-mails-weltweit/>
- [RFC5322] P.Resnick
Internet Message Protocol
Network Working Group, October 2008
- [Egli 01] Peter R. Egli
Introduction to mail transfer protocols for the internet, 2015
- [Tool 01] What is my IP address, last request: April 04, 2017
<http://whatismyipaddress.com/trace-email>
- [Inet 04] Patrick Föllner, December 09, 2014, last request: April 04, 2017
<https://www.seonative.de/unterschiedliche-ergebnisse-bei-gleicher-suchanfrage-wie-google-suchergebnisse-personalisiert/>
- [Tool 02] Tineye, last request: April 04, 2017
<https://www.tineye.com/>
- [Tool 03] CQCounter, last request: April 04, 2017
<http://cqcounter.com/whois/>
- [Inet 04] Jörg Breithut, September 15, 2016, last request: April 04, 2017
<http://www.spiegel.de/netzwelt/web/so-findet-facebook-ihre-freunde-a-1111698.html>

[Chung 01] Winnie Chung, John Paytner, Department of Management Science and Information Systems

Privacy Issues on the Internet, 2002

[Java 01] Adobe Acrobat SDK

Developing Acrobat Applications Using JavaScript

November, 2008

Image sources:

Figure 1: <http://www.statisticbrain.com/google-searches/>, last request April 23, 2017

Figure 3: <https://www.quora.com/What-is-the-algorithm-used-by-Google-reverse-image-search-i-e-search-by-image>, last request April 23, 2017

Figure 6: <https://www.divsi.de/wp-content/uploads/2014/07/canvas-abweichung.png>, last request April 23, 2017

Figure 10: <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>, last request April 23, 2017

Figure 15: <http://duff-johnson.com/2014/02/17/the-8-most-popular-document-formats-on-the-web/>, last request April 23, 2017

Curriculum Vitae



PERSONAL DETAILS

Name: Dusan Jaguzovic
Date of birth: 26.06.1990
Place of birth: Mrkonjic Grad, Bosnia and Herzegovina
Nationality: Bosnian and Herzegovina

EDUCATION

Since 06/2014 **Master of Science**
Networks and Security, JKU University, Austria

10/2010 – 06/2014 **Bachelor of Science**
Computer Science, JKU University, Austria

09/2005 – 06/2010 **High school**
Linz, Austria
Higher Technical Education Institute

09/1997 – 06/2004 **Elementary School**
Mrkonjic Grad, Bosnia and Herzegovina

PROFESSIONAL EXPERIENCE

Since 2/2015 **XXXLutz, Wels (Austria)**
Network Engineer

LANGUAGGE SKILLS

Serbian: Native
German: Fluent
English: Fluent
Spanish: Basic (A2)
Russian: Basic (A2)

INTERESTS

Cars, Motorbikes, Sports, IT Security, Mobile Applications, Web Applications

Linz, 25.05.2017

Signature