Assessing the Role of Online Banking Characteristics in the Target Selection of Banking Malware

Samuel Natalius

Master Thesis

August 2018

TU Delft

Assessing the Role of Online Banking Characteristics in the Target Selection of Banking Malware

by

Samuel Natalius

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday August 29, 2018 at 1:00 PM.

Student number:	4608380	
Project duration:	February 5, 2018 – August 29, 2018	
Thesis committee:	Prof. dr. ir. M. J. G. van Eeten,	TU Delft, chairman
	Dr. ir. C. H. Gañan,	TU Delft, 1^{st} supervisor
	Dr. ir. G. A. de Reuver,	TU Delft, 2^{nd} supervisor
	Dr. ir. S. Tajalizadehkhoob,	TU Delft, daily supervisor

This thesis is confidential and cannot be made public until August 29, 2018.

TU Delft

Preface

This thesis project is a final requirement for obtaining a master degree in Complex System Engineering and Management (CoSEM) in Delft University of Technology. The project, which has been ongoing for more than six months, is not only a requirement for completing my degree, but also an avenue for me to develop both my hard and soft skills as well as my personality in order to become a better researcher and a better person. There were lots of struggles and emotional moments I encountered during the completion of this thesis. Yet, with the help, support and guidance from those around me, I am able to maintain my spirit to keep on progressing and finally finish this thesis. I would like to express my gratitude to everyone who has supported me up to this point.

First and utmost, I would like to praise the Lord Jesus Christ for all His guidance in my life. I would also like to thank my beloved parents who have given me their everlasting support, especially during the hardest times.

I would like to thank all professors in my thesis committees, Prof.dr. M.J.G. (Michel) van Eeten, Dr.ir. G.A. (Mark) de Reuver, Dr.ir. C. (Carlos) Hernandez Ganan and Dr.ir. Samaneh Tajalizadehkhoob. A special thank to Michel as the Chairman for giving me a chance to research this topic. I would also like to express my highest gratitute to Samaneh, who first introduced me to the topic of this thesis. She did not only become my daily supervisor, but also my mentor and councellor during the difficult times I experienced in completing this thesis. I also appreciate any academic advice and guidance that my first supervisor, Carlos, gave to me, especially those related to statistical techniques. I would not be able to finish this thesis without all your support.

I also want to thank Fox IT who allowed me to use their data for this thesis. Also, thank you to Antonios who was willing to spend his time and effort responding to my questions especially those related to the data I used. My gratitude also goes to Mr. Paul Samwel, Mr. Maarten Jak and Mr. Huub Roem for their willingness to share their knowledge, experience and opinion in the interview sessions.

I also want to thank my cousins, Gaby and Pingkan, for helping me proofread this thesis. Also to all of my friends and colleagues in Delft University of Technology: All of my Indonesian friends (especially Darli and Rina, thank you for all your help and support), Akropoli group, all fellow CoSEM friends, Elsa (thank you for sharing the knowledge about data analysis and Python) and others.

Big appreciation also to my friends from my bachelor degree, especially Daniel, Ignatius and friends in Hese Paeh group. Thank you for sharing your knowledge and experience about cybersecurity. Thank you also to for friends in Marie van Jessekerk Delft for supporting me with your prayers. I also received significant help from many online tools, such as search engines, technical forums, open data platforms, journal databases etc. in completing this thesis. There are also many others who I cannot mention one by one, who have supported me in different ways. I would like to thank them as well.

Despite all effort spent on completing this thesis, it is still far from perfect. Yet, I hope that this thesis could bring a significant contribution to the research community and the society, and also inspire other researchers in their research in the field of cybersecurity.

Samuel Natalius
Delft, August 2018

Executive Summary

This thesis investigates target selection of the banking malware in the banking sector. Information Technology (IT) has been bringing huge improvements to the banking sector since the first appearance of Internet banking in the mid-1990s. Nowadays, banking malware receives great attention as it is capable of infecting numerous victims in a short time, which can consequently creates huge financial losses. The issue of banking malware, like many other cases in cybersecurity, is complex since such problem not only lies in the technical layer but also socio-technical and governance layer (van den Berg et al., 2014). There are multiple actors with multiple interests playing in the field and the landscape itself is also evolving over time.

Understanding target selection of the banking malware is a step before making a suitable proactive measure to address the issue. Therefore, several researchers have delved into this topic obtain better understanding of the target selection. Notwithstanding many good studies in the field, some gaps in the research are still present. The previous studies are more focused on the targeted entities, leaving the non-targeted entities unknown. Furthermore, some potential factors which influence the target selection may have not been addressed, and if so, may not have been assessed either. Malware data keeps updating, so research should also follow the updates.

Seeking to address the gaps, this research is conducted in order to find out **what characteristics related to online banking services can affect the likelihood of the malware attack** to them .

The study started with exploring the literature in order to collect characteristics which have the potential to explain the target selection. Quantitative analysis was performed afterwards. The research used malware attack data from February 2014 - November 2017 which was then linked to the list of registered banks in countries of the European Union (EU). Therefore, this thesis covers banks in the EU area in its scope. This thesis assesses some of the characteristics which potentially explain the target selection, namely the language (offered by online banking) and the authentication factor, hence the related data were collected by manual observation. Other data were also collected or extracted as controlling factors. Finally, expert interviews were conducted in order to gather perspectives about the target selection and also to comment on the model and analysis made in this thesis.

There are some interesting insights which were extracted from the quantitative analysis performed in this thesis:

- Out of 5,039 banks in the EU, 1,188 banks were without any online banking services and from 3,851 banks with an online banking service, 1,802 banks were found targeted and 2,049 not targeted. This indicates a lower number of targeted banks than the non-targeted.

- Aligned with a research by Tajalizadehkhoob (2013), power law distribution is still present in the malware attack, where a small percentage of banks (20 percent) attracted more than 80 percent of the attacks, regardless of the metrics formulated for counting the attacks.

- Different malware families were shown to target different banks. Some of them were seen to be used only for a specific attack, as indicated by the findings that some of malware variants only attacked one or a few countries and in a very specific time in the period, for example ReactorBot and PkyBot.

- Every country had a different ratio of the number of targeted banks versus the number of non-targeted banks. It is shown that in countries like the Poland and Ireland, only a small proportion of their banks were targeted, which can potentially be explained by the fact that cooperative banks were common in both countries. Meanwhile, countries like Croatia and Bulgaria have almost all of their banks targeted.

This thesis takes into account the argument from Tajalizadehkhoob (2013) that counting the attack rawly from the malware attack dataset may not result in the actual attack count as an update of the configuration file may cause the data of the same attack to be multiplied, thus, raising an overcounting problem. This thesis investigates the way the malware data is stored and infers that unique attack ID may be used to approach the actual attack count. Some metrics are proposed. A metric "week-interval attack count" takes the dimension of time to approach the actual attack count, assuming that new attacks with similar traits that occured before 7 days (a week) are updates to the primary attack. Another metric "Count of unique attack ID" was based on the investigation on the way the malware data is stored. The latter metric is seen to be the most reliable metric amongst all extracted metrics in approaching the actual attack count.

In addition, other metrics that can be extracted from the collected data were also presented. These were number of weeks a bank is under threat, number of unique URLs corresponding to the bank, number of different malware variants targeting an online banking, the presence of a particular language, the presence of authentication factor, and the ranking of an online banking domain.

The explanatory analysis was conducted to see whether there are relationships between external factors, in particular languages and authentication factor, and the attack count. Two kinds of statistical models were created: logistic regression model, to see whether the factors could explain the probability that an online banking entity will be targeted or not targeted, and and negative binomial regression model, to see whether the factors could explain the tendency of the online banking entity being more or less targeted among the targeted entities. The negative binomial regression model was made for every attack count metric.

Despite the presence of control variables in the models like the country and domain popularity, some languages are shown to have more significance than others in explaining whether an online banking entity is more or less targeted. The presence of two-factor authentication (2FA) is shown to have significance to reducing the tendency of an online banking entity to be less targeted. The research also showed that these characteristics still maintain their significance despite the presence of control variables. However, looking at several factors which became less significant as more control variables were included in the model, there is an indication that some factors may be more important than others in terms of affecting the target selection.

Next, interviews with experts' were conducted. The purpose of the interviews was to get experts' perspectives on target selection and factors they believe could influence the target selection as well as to get their opinion and interpretation of the models. It is inferred from the interviews that, even though experts found the models rational, additional factors, especially financial- and market-related factors, could be added to the models in order to make a more plausible conclusion about target selection out of them.

In conclusion, this research shows that there are several characteristics that potentially explain the likelihood of an online banking entity being attacked by the malware. However, enhancing the current model by adding more factors is required to make the conclusion more convincing. This leads to some limitations and suggestions that may be useful for future research.

Contents

List of Figures

List of Tables

Abbreviations

AIC	Akaike Information Criterion
AUC	Area Under Curve
C&C	Command and Control
CSV	Comma-separated Value
EBA	European Banking Authority
ECB	European Central Bank
EU	European Union
FTP	File Transfer Protocol
GDP	Gross Domestic Product
GLM	Generalized Linear Model
HTML	Hypertext Markup Language
HQ	Headquarter
IT	Information Technology
P2P	Peer-to-peer
POP3	Post Office Protocol, version 3
RAT	Routine Activity Theory
RCT	Rational Choice Theory
ROC	Receiver Operating Characteristic
SAT	Situational Action Theory
SQL	Standard Query Language
TLD	Top-level Domain
URL	Uniform Resource Locator

1 Introduction

1.1. Background

Our world has come to a state where information technology (IT) has become an integral part of many aspects of life. The intense penetration of personal computer and mobile devices along with the development of Internet causes modern society to depend on IT in their lifestyle. IT also changes the way businesses are performed in many sectors, including the financial sector.

IT brings huge improvements to the financial sector. Take banking as an example, which has realised the value of using electronic platforms as efficient channels to reach more customers and performed transactions with less effort and cost since the first appearance of Internet banking in the mid-1990s (Jaleshgari, 1999; Vrancianu & Popa, 2010). The product range, product development, service channels, service type and packaging in the banking sector has been transformed by IT development (Campanella, Della Peruta, & Del Giudice, 2017).

On the other hand, IT in the financial sector also raises new types of issues, especially those related to security. Besides physical- and human-related security issues, users and facilitators should also consider digital-related security issues. Moreover, the physical, digital and human aspects may also fuse in the cyberspace, causing a need to address the issue in a more holistic way. Financial sector is a hot target for attacks due to it having many valuable assets, and nowadays its reliance on IT makes it prone to various cyberattacks (PlugandPlay Tech Center, 2017). To illustrate, cybercrime accounts for almost USD 114 billion of financial losses in banking sector globally (Raghavan & Parthiban, 2014). Amongst all possible types of cyberattacks on the financial sector, the malware-related cyberattacks are interesting to delve into since many attacks are based on the financial malware (Tajalizadehkhoob, Asghari, Gañán, & van Eeten, 2014), especially considering its potential to infect a large number of entities and create a huge loss.

Analysing malware attacks in the financial sector cannot be done straightforwardly. Aligned with the concept of cyberspace by van den Berg et al. (2014), this problem not only lies in the technical layer but also socio-technical and governance layer, implying that there are multiple actors, with their own interests, values and goals, playing on different aspects in this field, thus making this problem complex. Any changes in a cyberspace layer can influence other layers. For example, stakeholders' perspective (socio-technical) can influence how they regulate their security (governance) which alters the technical aspects of the protection. Its complexity is also intensified with the dynamicity of the sector. Financial sector is constantly developing in terms of its service trends, customer expectations, business model, etc. (Craig, 2016). The cyberspace is also dynamic and the cyberattacks are also evolving over time.

1.2. State-of-the-art Research

This section explains the context of the thesis topic and the latest research on cybersecurity in the financial sector, which are useful for identifying the knowledge gaps and the main research question.

1.2.1. Information Technology in Financial Sector

IT has played a role in financial sector since the 1990s when banks started to utilise Internet to access their computer systems (Claessens, Dem, De Cock, Preneel, & Vandewalle, 2002). Since then, digitalisation has played a huge role in transforming the financial sector. The term "Digital Finance" refers to this digitalisation in general (Gomber, Koch, & Siering, 2017). Digital finance covers any kind of electronic products and services of the financial industry, from credit and chip cards to home trading services, from Automated Teller Machines (ATMs) to mobile and app services (Banks, 2001; Gomber et al., 2017).

Digitalisation not only affects the technical aspect, but also creates a social impact. The view on IT in the financial sector is currently shifting, from treating it as a tool to seeing it as a business opportunity. It is common nowadays, for example, for people to easily access their bank account and perform financial transactions via their PC, laptop or even mobile phone as if they are physically going to the office of the bank they use.

1.2.2. Development of Financial Malware

Due to the development of IT in financial sector, many financial firms, especially banks, and the society become strongly dependent on IT to perform financial transactions. While many advantages can be seen from this trend, for example, higher availability rate and efficiency of the services, there are also new types of risks arising related to cyberattacks.

Amongst many types of financial cyberattacks, financial malware is getting more attention due to its capability to infect numerous victims in a short time. For example, one of the most popular banking trojan, Zeus, was said to be "The king of malware" as it has been accused for causing the highest infection rates and financial losses (IZOOlogic, 2016). Financial malware also evolves very rapidly. Zeus transformed its business model from "freemium" to "as-a-service" after its source code leaked in 2011 (Bottazzi & Me, 2014; Hutchings & Clayton, 2017). Consequently, new types of malware also appeared; some of them were derived from Zeus. Citadel, for example, is known as one of the most successful financial malware derived from Zeus after its source code was leaked (Etaher, Weir, & Alazab, 2015). Many other variants are also responsible for causing harms to the financial services, namely Gozi, Kronos, Qadars, Ramnit, etc.

One should put in mind that the malware is only a tool for performing cyberattacks (Bottazzi & Me, 2014; Hutchings & Clayton, 2017; Tajalizadehkhoob et al., 2014; Wyke, 2011). Its attack decision, however, lies on the attack strategy and the behaviour of the attackers. In order to analyse cyber risks within the financial sector, one should consider the technical context, the socio-technical context and the governance context altogether (van den Berg et al., 2014).

1.2.3. Research on Target Selection

Potential victims may defend the attacks in either a proactive or reactive manner. A proactive approach seeks to understand risks before an event happens while a reactive approach tries to detect and stop an attack as it occurs. (Xu, Bailey, Vander Weele, & Jahanian, 2010). Proactive defence can arguably reduce the negative impact of attacks more significantly than reactive defence since it makes the system more dynamic and harder to predict (Lu, Marvel, & Wang, 2015). Asghari et al. (2016) argued that incentives shape the behaviour of actors in the cyberspace including attackers. For example, attackers are likely to select targets and attack strategies based on their expected financial/political benefits and risks. This argument can be a basis of performing a proactive measure to the attacks. Identifying the incentives might not be straightforward, but it could be approached by understanding the target selection.

Target selection was once defined as attack choices by financial cyber criminals utilising malware as to which financial institution to attack, at which point of time and for how many weeks (Cheung, 2017; Van Moorsel, 2016). Van Moorsel (2016) argued that having a clear understanding of the threat landscape is essential to further develop cyber risk management. In the case of online banking, for example, the attacks were becoming more target-specific, requiring a conscious selection process on the criminals' side (Tajalizadehkhoob et al., 2014; Sherstobitoff, 2012; TrendMicro, 2012). Findings from the target selection analysis will be beneficial for designing effective risk management.

There are some researchers who have delved into target selection analysis of financial malware. Tajalizadehkhoob et. al. (2013; 2014) set a basis of target selection research: extracting intelligence on criminal attack patterns and target selection from Zeus configuration files, identifying services targeted by Zeus malware, factors that could explain target selection, effects of the source code leakage to the target selection, etc. Another research compared the experts' view and expectation on target selection with the insight from the analysis on Zeus malware dataset (Van Moorsel, 2016). Cheung (2017) also analysed factors affecting the target selection of distributed denial of service (DDos) attacks on financial services. Another research by Tajalizadehkhoob et. al. (2017) assessed the role of hosting providers in combating financial malware's command and control (C&C) infrastructure, using their previous research as reference. Finally, Hutchings & Clayton (Hutchings & Clayton, 2017) researched the topic in a more qualitative way: analysing online cybercrime forum posts discussing Zeus configuration to find whether the discussion has evolved with market conditions and externalities.

In general, lots of previous studies have been done on understanding the target selection of Zeus malware, whereas the targets were mainly specified to banks, especially their online banking. The analysis of the target selection was mostly focused on assessing the targets' general characteristics, for example the size and type of the bank.

1.2.4. Knowledge Gap and Research Question

The previous research on target selection has set a strong basis in analysing how likely financial institutions, especially banks, are being targeted. Yet, there are still some knowledge gaps in the current state of the research. First, the previous studies focused more on the institutions that were targeted. On the other hand, not much was known about those who were not targeted. The insights from the non-targeted one therefore could add the current knowledge about the target selection. Second, some potential characteristics which influence the target selection were not yet addressed, and if so, were not yet assessed either, for example the authentication factor. Lastly, the malware data also keeps updating and so research should also follow the updates. So far, it is seen that the previous studies limited their scope to one type of financial malware family, i.e. Zeus, and some of them used data which were from more than 5 years ago. Having the latest insights will definitely bring the most recent understanding about the development of this issue.

Seeking to bridge these knowledge gap, the thesis tries to observe factors related to online banking services that potentially affect the likelihood of an online banking entity being targeted by malware criminals. For this purpose, the following main research question (MRQ) is made:

What characteristics related to online banking services can affect the likelihood of the malware attack?

Due to the availability of data, the research will consider the banks registered in the selected European Union (EU) countries. The attack data used for this research came from the attack database of Fox IT, a leading cyber security company in the Netherlands.

The result of this thesis would be very useful for banks, as the potential victims of this situation, to re-evaluate their position in the malware threat landscape and make decisions for their proactive measures.

1.3. Research Approach and Methodology

A quantitative analysis was used to approach the main research question (MRQ) of this thesis. A quantitative research uses data and statistical procedures to examine the relationship between variables (Creswel, 2008), in this case was the relationship between characteristics of online banking services and the likelihood of malware attack. However, before the analysis can be conducted, it is important to identify the characteristics that could potentially influence the target selection of financial malware. This leads to the first research question:

> *RQ1. What characteristics of banks or their online banking services can potentially explain the likelihood of them to be targeted?*

In order to answer RQ1, scientific and professional literature was collected and analysed. It was done in order to obtain updated information, theories and frameworks necessary for identifying the characteristics of the banks which potentially influence the target selection of financial malware.

Some of the identified characteristics were selected for further analysis. To enable the analysis, these characteristics were further derived into quantifiable variables and metrics, along with the identification of any possible confounding variables which might be present in certain characteristics. Later, variables and metrics were selected by counting several considerations, such as the relevance of the variables and metrics in justifying the characteristics assessed and the availability of data for the calculations. The result from this study became the basis for the next phase. RQ2 is thus formulated to summarise this process:

> *RQ2. What quantifiable metrics and variables relevant to the characteristics can be collected for the analysis?*

Before going further to the analysis of the selected characteristics which are manifested in the metrics and variablef from RQ2, it is beneficial to understand at the data themselves in order to get a sense of the landscape of the attack target as presented in the data. To accommodate this purpose, RQ3 is formulated as below:

> *RQ3. How does the target landscape of online banks look like?*

Once feasible variables and metrics were established, a quantitative study was performed for testing whether the factors in RQ1, as reflected in the variables and metrics, have significant effect to the target selection. This research question is formulated for this point:

> *RQ4. Which of the selected characteristics could explain the extent the online banking were targeted by the malware?*

The thesis applied the following methodology to answer the above research questions:

1. The research started by performing literature review for defining general characteristics and security measures which will be taken for the analysis together with the justification of why such characteristics could explain the target selection of the malware. Literature review is deemed necessary as those elements should be supported by strong theoretical argumentation that justifies their relevance to the issue. The result of this stage is a conceptual framework for answering RQ1.

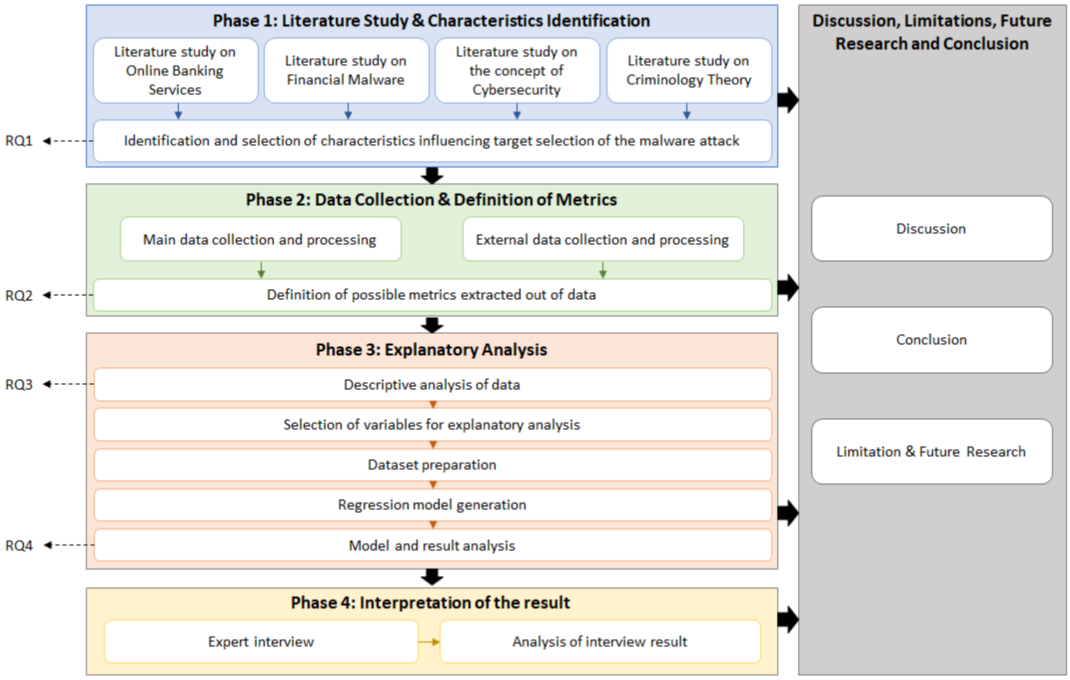

Phase 1: Literature Study & Characteristics Identification
Literature study on Online Banking Services
Literature study on Financial Malware
Literature study on the concept of Cybersecurity
Literature study on Criminology Theory
RQ1
Identification and selection of characteristics influencing target selection of the malware attack
Phase 2: Data Collection & Definition of Metrics
Main data collection and processing
External data collection and processing
RQ2
Definition of possible metrics extracted out of data
Phase 3: Explanatory Analysis
RQ3
Descriptive analysis of data
Selection of variables for explanatory analysis
Dataset preparation
Regression model generation
RQ4
Model and result analysis
Phase 4: Interpretation of the result
Expert interview
Analysis of interview result
Discussion, Limitations, Future Research and Conclusion
Discussion
Conclusion
Limitation & Future Research

Figure 1.1: Research flow diagram of this thesis

2. Exploration of the collected data as well as obtaining information from literature and specialists were done in order to find the necessary quantifiable factors, variables and metrics of selected characteristics, addressing RQ2.

3. Data analysis were then performed for evaluating whether the characteristics, as defined in RQ1, significantly explain the pattern from the data analysis, using variables and metrics from RQ2. The analysis began with descriptive analysis in order to get much understanding about the target landscape of online banks, answering RQ3. Furthermore, exploratory analysis were conducted to see the relationship between the characteristics that were selected for the assessment and the target selection, addressing RQ4 of this thesis.

4. Finally, expert interviews were conducted for obtaining their opinion about the target selection of financial malware and gathering their critical review and interpretation of the result of the quantitative analysis.

Research Flow Diagram (RFD) as in figure 1.1 provide an illustration on how the thesis is conducted in accordance with the research methodology explained previously. There are four main phases, each phase corresponds to a research question of this thesis. In the end, the discussion, arguments about the limitation of the research, possibilities for future research and conclusion will be provided.

1.4. Outline of the Thesis

The outline of this thesis is as below:

- Chapter 2 will provide the review of the current literature about the concept of cyberspace, cybercrime and economics of cybersecurity, the concept, mechanism and security of online banking services, the overview of financial malware, the actors in the financial malware attack and theory from criminology field in order to analyse the characteristics influencing possible target selection. The last part of the chapter will provide the identification of potential characteristics based on the literature study.

- Chapter 3 will explain the process of collecting and preparing the datasets which will be used for the analysis.
- Chapter 4 will present metrics which can be extracted from the collected data, including metrics that will be used to assess the selected characteristics.
- Chapter 5 will provide the descriptive analysis of the collected data, so that a better understanding of the data as well as interesting insights about the target landscape can be obtained.
- Chapter 6 will contain the explanatory analysis. The statistical model based on the collected data and the extracted metrics will be presented and the analysis following the model result will be given.
- Chapter 7 will provide the outcome of interviews that were conducted with the security experts in the banking sector.
- Finally, chapter 8 will conclude the thesis by discussing the performed study, stating the contribution of the thesis and presenting the limitation and future research.

2 Literature Review and Selection of Characteristics

2.1. The Concept of Cyberspace and Cybercrime

Before going deeper to the cyber attack case in online banking, it is important to first understand cyberspace and cybercrime, as they are two fundamental concepts related to the situation explained in this thesis.

Singer & Friedman (2014) defined the cyberspace as "the realm of computer networks (and the clients behind them) in which information is stored, shared, and communicated online" (p. 13). This definition emphasises technology, in this case is computer networks, as the key for the cyberspace to exist. Although the definition is true, it is rather limited in describing the complexity of the cyberspace known today. Van den Berg et. al. (2014) argued that the human and society factors should also be considered as well and thus proposed two other layers, namely the socio-technical layer and the governance layer, on top of the technology layer in the concept of cyberspace. Figure 2.1 illustrates the concept of cyberspace according to van den Berg et. al. (2014). It is seen that they also divide the cyberspace into many sub-sectors where one of them is the financial sector.

Cybercrime is a type of activity which, according to van den Berg et. al. (2014), is enabled in the socio-technical layer of the cyberspace. There are many sources which try to define the cybercrime (Nagurney, 2015; Petee, Corzine, Huff-Corzine, Clifford, & Weaver, 2010; Raghavan & Parthiban, 2014; Yar, 2005; Thomas & Loader, 2000; Bougaardt & Kyobe, 2011; Brenner, 2006; Gordon & Ford, 2006; Cheung, 2017). By extracting these sources, a definition of cybercrime is formulated for this thesis: any activity that is illegal or considered illicit by certain parties, committed or facilitated by one or several entities using the capabilities of the electronic, computer and network technology, which may cause disadvantages to certain parties or the society. Cybercrime should be treated seriously by financial firms. According to the Global Economic Crime and Fraud Survey (PwC, 2018), cybercrime accounts for 41 percent of the reported frauds in the financial services, placing it second on the rank of the most reported frauds in the sector after consumer fraud, together with asset misappropriation. The fraction of cybercrime in the financial services is also the highest among other sectors assessed by the report. Therefore, cybercrime is one of the most significant threats to financial firms' business continuation.

2.1.1. Economic Cost of Cybercrime

The (economic) cost of cybercrime is, although enormous, hard to quantify. However, it can be systematically analysed by categorisation. Combining the arguments of Anderson et al. (2013) and Lagazio et al. (2014), the economic cost can be grouped into four categories:

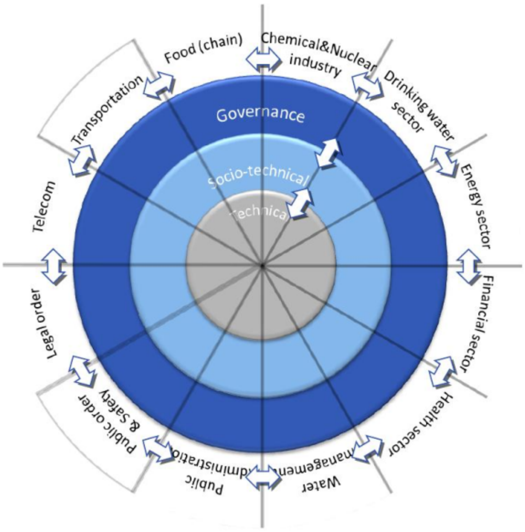


Figure 2.1: The Conceptualization of Cyberspace according to Van den Berg et. al. (2014)

- Criminal revenue, which is the gain received by the criminal from a crime.
- Direct loss, which is loss, damage, or other suffering experienced by the victim because of a cybercrime. For example, asset loss or compensation given to the unlucky clients.
- Indirect loss, which is loss or opportunity cost felt by the victim due to the existence of a cybercrime, whether it is successful or not. For example, reputational cost and loss of trust by the customers.
- Defence cost, which can be further grouped into:
  - direct defence cost for the development, deployment and maintenance of its countermeasures (for example, the annual cost of implementing antivirus), and
  - indirect defence costs caused by the negative consequences and opportunity costs caused by the implemented measures (for example, the cost of employee training and change management).

2.2. Online Services of Banks and Their Problems regarding Financial Malware Attack

2.2.1. Background of Online Banking Services

Internet banking emerged in the mid-1990 when many banks began to realise the potential of the Internet as a novel, cost-effective banking channel (Calisir & Gumussoy, 2008; Claessens et al., 2002; Gopalakrishnan, Wischnevsky, & Damanpour, 2003). It started with the Internet-enabled home banking software, but banks later began to provide the service through World Wide Web (WWW) as its development, maintenance and distribution were easier (S. Z. Kiljan, 2017). Nowadays, nearly all banks offer various financial services online and facilitate their clients accesses to their online services (Tajalizadehkhoob, 2013). Meanwhile, mobile banking, together with related payment technologies, are increasingly used following the market growth of Internet-enabled smartphones (Nand, Astya, & Singh, 2015). Mobile banking is considered more location independent than home banking (S. Z. Kiljan, 2017). Kiljan (2017) also argued that most of financial transactions are now performed through mobile banking and predicted that the number of mobile banking users is expected to increase exponentially in the period 2020-2025.

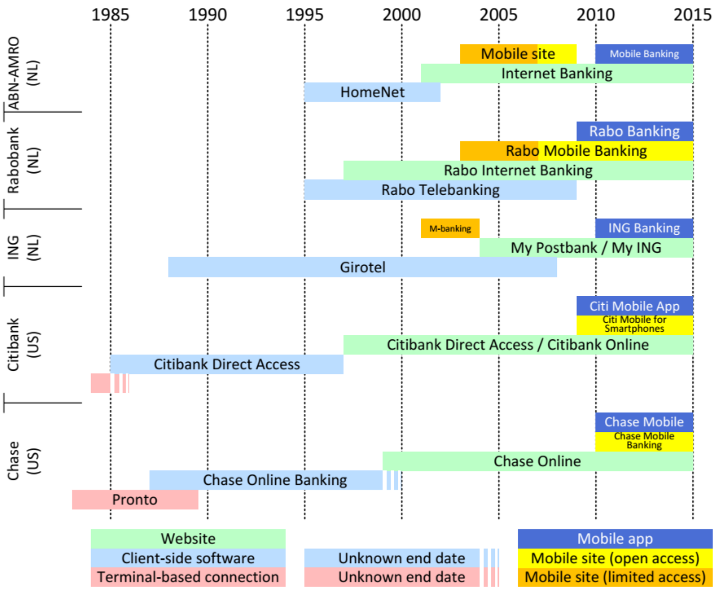


Figure 2.2: The development of online banking in several banks in the United States (US) and the Netherlands (NL), taken from (S. Z. Kiljan, 2017)

In terms of the development, United States (US) and the Netherlands (NL) were among the countries that adopted online banking services early. Figure 2.2 illustrates the development of online banking in several banks in US and NL in the period 1985-2015. It shows that nowadays mobile service is offered to complement the web service.

2.2.2. The Mechanism and Security of Online Banking

Extracted from several sources (Tajalizadehkhoob, 2013; Nand et al., 2015; Van Moorsel, 2016; Hutchinson & Warren, 2003), a current online banking service, in general, comprises these main components:

1. The client side, consisting of services and applications that enable customers to access their bank accounts and perform financial transactions. It can be in many forms, for example, the e-banking website intended to be accessed through personal computers (PCs) and laptops, the mobile banking site (m-banking) which is a light website designed to be accessed from mobile browsers and the mobile application in smartphones. Internet access is required in order to make such services and applications work properly.

2. The network infrastructure, which is required in order to connect the client-side services and applications to those of the bank-side. Nowadays, Internet is the most widely used medium.

3. The bank side, more specifically, its web server and backend system. The web server serves and processes requests coming from the client side. It is linked to the backend system which holds the main centralized database storing all the crucial data related to the bank and its customers.

Figure 2.3 illustrates the above components and their relationships.

Regardless of the type of devices used to perform the transactions, both home banking (accessing the bank's services through PC or laptop) and mobile banking (accessing the services through

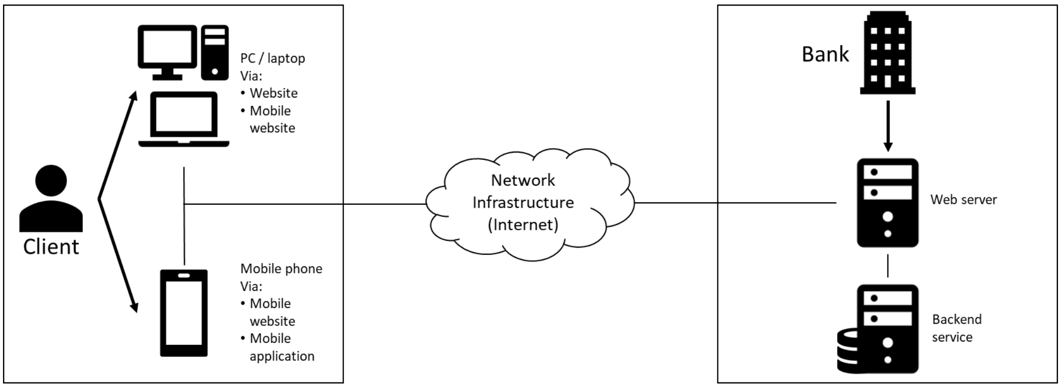


Figure 2.3: The components of online banking and their relationships

mobile phones) are prone to cyberattacks as long as it is connected to the Internet. However, Kiljan (2017) claimed that, until now, mobile banking is relatively safe due to the following factors:

- Mobile banking has yet to reach home banking's level of popularity.

- Functions in mobile banking are generally fewer than those in home banking and they overlap with the home banking functions.

- Since almost all home banking users access the services through websites, malware developers only need to write the malware once and easily customise it for each targeted bank site. However, mobile malware is harder to customise as there are many mobile operating systems which forces banks to write codes specific to each system.

Therefore, assessing cyberattacks on the home banking websites is still relevant in the present times. Hence, this thesis will focus on the malware attack on the home banking, particularly the online banking website.

To defend against cyberattack, banks implement some security measures to their online banking services. One of the measures is the authentication process. In general, it is a process done so that "one agent should become sure of the identity of the other" (Lowe, 1997). The authentication process can vary in terms of its methods and the level of security. The methods may range from the simple static username and password to two- or multi-factor authentications combining two or multiple authentication factors, such as something the user knows (knowledge), something the user physically has (possession) and something the user physically is or does (biometrics) (Claessens et al., 2002; S. Z. Kiljan, 2017; Tajalizadehkhoob, 2013). So far, there are two forms of authentication for authorising financial transactions in online banking (S. Kiljan, Simoens, De Cock, van Eekelen, & Vranken, 2014; S. Kiljan, Vranken, & van Eekelen, 2018):

- Entity authentication aiming to prove the identity of an online banking user, and

- Transaction authentication aiming to prove that a certain transaction is intentionally performed and authorised by the right user.

2.3. Financial Malware in Focus

This section will explain the profiles of malware variants which have major presence in the attack target dataset used in this thesis.

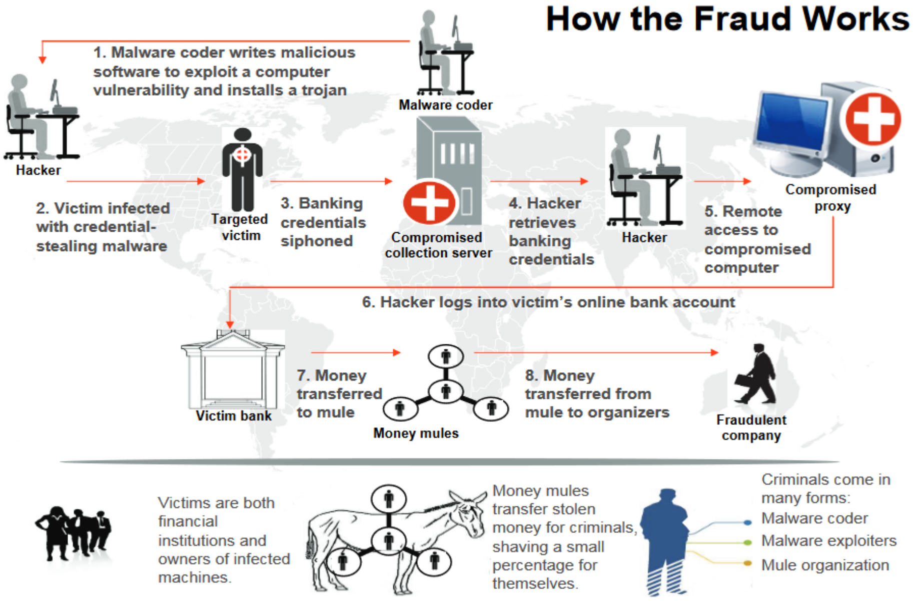

How the Fraud Works
1. Malware coder writes malicious software to exploit a computer vulnerability and installs a trojan
Malware coder
Hacker
2. Victim infected with credential-stealing malware
Targeted victim
3. Banking credentials siphoned
Compromised collection server
4. Hacker retrieves banking credentials
Hacker
5. Remote access to compromised computer
Compromised proxy
6. Hacker logs into victim's online bank account
Victim bank
7. Money transferred to mule
Money mules
8. Money transferred from mule to organizers
Fraudulent company
Victims are both financial institutions and owners of infected machines.
Money mules transfer stolen money for criminals, shaving a small percentage for themselves.
Criminals come in many forms:
Malware coder
Malware exploiters
Mule organization

Figure 2.4: How the Fraud Works (Federal Bureau of Investigation, 2010)

2.3.1. Zeus Malware

Zeus is referred to as the "king of bank fraud trojan viruses" (Infosecurity Magazine, 2010). Zeus, also known as Zbot, WSNPOEM, NTOS, or PRG, is a trojan horse running mostly on Microsoft Windows environment (also Blackberry and Android phones since 2012), capable of performing man-in-the-browser (MitB) attack to its victim and stealing credentials (Tajalizadehkhoob, 2013; Van Moorsel, 2016; Etaher et al., 2015). It is essentially a malware kit containing tools to build and control a botnet which is so simple to use that it is massively traded in underground media by cybercriminals (Tajalizadehkhoob, 2013; Wyke, 2011). It is also the basis of some other financial malware, like ICE IX, KINS and Citadel, especially after its source code leaked in May 2011 (Etaher et al., 2015).

According to Falliere & Chien (2009) from Symantec, there are four main actions performed by Zeus:

- gathering and sending of system information to the command and control (C&C) server. Some of the details gathered include version of the bot, operating system (OS) version, OS language, the country of the compromised computer, and its IP address,
- Stealing of information stored in PSTORE (Protected Storage) as well as FTP passwords and POP3 passwords,
- Stealing of online credential information as specified by a configuration file. It is mainly done by web page injection, that is, injecting additional HTML to legitimate pages which tricks users to provide private information out of what is required by the website, and
- Communication with the C&C server for the next action. The next action can be rebooting or shutting down the computer, deleting the computer's system files, initiating back door, etc.

An infographic from the Federal Bureau of Investigation (2010), as seen in Figure 2.4, depicts the process of online banking frauds using Zeus (and probably its derived variants). There are 4 main processes which will be explained below.

1. Malware coding and customising

This process covers the first step of the flow in the infographic. It is clear that the malware must first be designed, coded and built for it to work. As for Zeus, it is prepared as a kit which is traded among cybercriminals. Therefore, offenders do not have to code the malware from scratch, but simply customise it to suit their specific purposes. Every Zeus kit consists of a control panel application for maintaining/updating the botnet as well as retrieving/organising recovered information, and an EXE builder for creating the Trojan binaries and encrypt the configuration file (Tajalizadehkhoob, 2013).

The configuration file of Zeus malware is responsible for providing the initial configuration of the malware and hence, it is the file that is customised by the attacker before the bot executable is generated. The configuration file consists of two parts (Falliere & Chien, 2009):

- the static configuration, containing parameters that are not changed during the life of the generated bot executable, and

- the dynamic configuration, containing parameters which will be updated regularly. This part will generate a dynamic configuration file (config.bin), encrypted with a key. When the trojan is active, it will download the generated file at regular intervals.

Once the configuration file is established, the attacker will need to build his/her own bot executable using the builder available in the kit. This bot executable will then be distributed to the victims.

2. Infection process

This process covers the second step of the flow in the infographic. The infection happens when the victims intentionally or accidentally execute the infected file(s). The infected file(s) can be spread in multiple ways, such as by clicking on a spam email, downloading it from a malicious website, getting it from external storage media, etc. A computer infected by the trojan will become a bot, and it will initiate the communication with the C&C server and ask for a dynamic configuration file, which has been generated previously (Tajalizadehkhoob, 2013). It will also retrieve the URL of the drop server; the server which allow the bot master to monitor the bot's status, issue commands to the bot and retrieve the information collected by the bot, including the stolen data (Tajalizadehkhoob, 2013).

3. Credential stealing

This comprises the third and the fourth step of the flow in the infographic. Credential stealing starts by instructing the bot (from the C&C server) to send commands or inject codes into the bank webpage loaded by the browser of the infected computer (Tajalizadehkhoob, 2013). The injected codes mostly contain text fields to get the input of credential information from the victims. The unaware victims will probably enter the asked credential information thinking that the website is legitimate without knowing that such information will be sent to the attacker. It is a form of man-in-the-browser (MitB) attack.

Once the credential information is entered, it will be sent to the corresponding drop server. The attacker will eventually receive the stolen credential from that server.

4. Money stealing

This comprises the fifth to the eighth step of the flow in the infographic. Once the credentials are acquired, the offender can remotely access the compromised computer to log into the victim's online bank account. Once logged in, the offender may perform the transfer of money to the money mule's account.

Zeus malware also evolved into several variants, most notably Zeus Action, Zeus P2P, Zeus OpenSSL and Zeus Panda. Zeus P2P, also known as GameOver Zeus (GOZ), is capable of running on top of a decentralised network of compromised computers, eliminating its dependency of a centralised C&C server and hence giving it more resilience against takedown actions (Andriesse, Rossow, Stone-Gross, Plohmann, & Bos, 2013). Zeus Panda is another variant which has a unique characteristic, in that it will become dormant when it detects the presence of any virtual machine or tool often used by malware, making it difficult to be detected (Berghoff, 2017).

2.3.2. Citadel

Citadel is a new variant of Zeus malware emerging after Zeus source code leaked in May 2011. At a glance, Citadel performs similarly to the original Zeus malware; however, some enhancements attract criminals to Citadel (Wyke, 2012). The enhancements are:

- The presence of new capabilities, such as video capture and DNS redirection.
- A built-in Customer Relationship Management (CRM) system to handle any issue quickly and effectively.
- Improved encryption techniques for the communication between the malware and its C&C server by applying Advanced Encryption Standard (AES) and, even in the latest version, another XOR layer on top of it.

Another notable feature of Citadel is the function `webinjects_update`, which provides a dynamic way of checking whether or not, and setting the way for how the malware performs the web injection on its victims' computers (Wyke, 2012).

Many of Citadel's victims are in Europe, especially in Germany, the Netherlands, and the United Kingdom (Tajalizadehkhoob, 2013). Tajalizadehkhoob (2013) also mentioned that its victims are not only limited to banks, but also other entities like government agencies.

2.3.3. Dridex

Dridex is another banking malware which mainly targets customers of Europe's major online banking institutions (Sanghavi, 2015). It works in a similar way as Zeus and its derivation, which is, by injecting HTML codes to manipulate the victims in order to steal their personal information.

First of all, the criminals try to infect a victim's machine by sending spams with their malicious attachments (usually macro-enabled MS Office files). The malware is installed and executed when the victims open the file and enable the macro for it. Once installed and executed, the attackers can then gain control of the victim's machine and perform many actions, e.g. upload/download files, take screenshots, add the compromised computer to a botnet, communicate with other peer nodes via peer-to-peer (P2P) protocol and inject HTML codes into browser processes to monitor communications and steal information.

According to Kaspersky Lab (2017), based on the malware activity in the early months of 2017, the UK accounted for nearly 60% of all detections, followed by Germany and France. Interestingly, the malware never works in Russia.

2.3.4. Ramnit

Ramnit is another banking malware which was first detected in 2010. It has the ability to steal cookies to hijack online sessions for banking and social media websites, steal FTP login credentials, inject code into web pages when the victim accesses certain sites, enabling attackers to remotely access the computer and steal files (Symantec, n.d.).

According to Symantec (n.d.), Ramnit mainly targeted users in Asia, with India being the most threatened followed by Indonesia and Vietnam. In contrast, the presence of Ramnit in Europe is small. Figure 2.5 shows the geographic distribution of Ramnit.

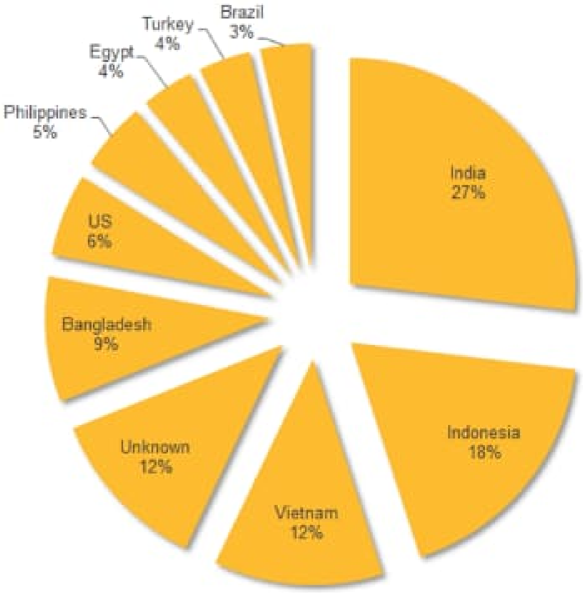


Figure 2.5: Geographic distribution of Ramnit (Symantec, n.d.)

2.3.5. Kronos

Kronos was first advertised around June 2014 by an individual who wrote the advertisement in Russian (Malwarebytes Labs, 2017a). This malware is quite unique, as it was often spread by various exploit kits and often used as the downloader of other malware.

Although Kronos is known as a banking malware, it not only targets banks but also several other popular sites (Malwarebytes Labs, 2017b). Some examples include Facebook, Citibank and Wells Fargo. Kronos steal its victims' data by doing phising, that is, injecting a HTML code which opens an additional pop-up form then asking the victims to enter their credential information there (Malwarebytes Labs, 2017b).

2.3.6. Other Banking Malware

There are several other banking malware variants which are present in the banking landscape. Many of them were variants of or coded on ZeuS malware. Qadars, for example, was created from the leaked source code of Carberp and Zeus. It targeted only several countries in Europe like the Netherlands, France and Italy and interestingly, is capable of bypassing mobile-based two-factor authentication of online banking (Boutin, 2013; Cimpanu, 2016). Another example, KINS, initially attacked financial institutions in Europe, specifically Germany and the Netherlands (Fox-IT, 2013).

Other banking malware also often share similar characteristics, such as having the capability to inject code to the browser to steal credential information. Gozi, for example, has been around since 2007 and, like ZeuS, also performs HTML injection to steal information (Ducklin, 2016). Another example, Timba, also performs web injection in order to steal personal information from its victims, mainly bank customers in Europe, especially Poland (Bach, 2015).

In conclusion, most banking malware uses web injection to steal credential information from its victims. Many of the malware variants are apparently developed based on the preceding malware, mostly ZeuS. However, their targets can be very diverse in scope; while some variants were known to target banks accross the world, some others might focus only at several countries or a specific region.

2.4. Actors in the Financial Malware Attack

Reflecting on the financial malware attack situation and condition, as well as looking at the general behaviour of the malware itself, it is possible to extract information of actors who are either directly or indirectly involved in the financial malware attack. Actors directly involved in it are victims, malicious actors and money mules. Security guardians, IT/network infrastructure service providers and authorities may be indirectly involved in the case of financial malware attack as well.

Victims

Victims of the financial malware attack cases are bank customers, which can be divided into individual customers, small-medium enterprises (SMEs) and large organisations/companies, and the banks themselves (Tajalizadehkhoob, 2013). Bank customers are likely to be directly affected by the attack as their assets (i.e. money, credentials) are prone to being stolen (OECD, 2008). Meanwhile, banks also become the victims as a result of both direct and indirect costs, such as asset loss due to giving compensations, reputation loss, loss of trust, legal consequences, etc.

Tajalizadehkhoob (2013) argued that individual and SME customers create the most negative externalities among the legitimate actors due to several factors:

- they often do not understand how malware works and whether they become infected,
- they perceive that they are unlikely to be attacked,
- they do not find paying for security software convenient, and
- they are not familiar with IT and find the Internet security matters difficult, making them the weakest link in cyber security environment (Asghari, 2010; Mannan & van Oorschot, 2008).

Malicious actors

OECD (2008) categorised malicious actors into 5 (five) categories:

- The Innovators: the malicious actors who, triggered by challenge and curiosity, actively try to find security holes in systems or environments and other new opportunities for exploitation.
- The Amateur-fame Seeker: the malicious actors who have limited competences in computing and programming skills but desire fame. They use established tools and procedures to perform their attacks.
- The Copycat-ers: the malicious actors who imitate proven simple attacks with the aim to gain more popularity in the cybercrime community.
- The Insiders: the malicious actors who has the knowledge of an entity's security quality thanks to their previous privileges and often perform the attack with theft or revenge as an aim.
- (Organised) Criminals: the malicious actors who often perform the attack with the aim to gain a profit. They are usually highly motivated, organised, knowledgeable and powerful.

In the context of financial malware attacks, the malicious actors can also be categorised into malware developers and malware users. Malware developers are actors who develop or improve the capabilities of the malware, while malware users use the developed or improved malware and, if applicable, customise it in order to perform the attack. Malware developers may develop or improve the malware to use it for their own purposes or to sell it to potential users in the (underground) market. This situation is evident in many financial malware cases, especially Zeus.

Money mules

Money mules are "individuals recruited wittingly and often unwittingly by criminals, to facilitate illegal funds transfers from bank accounts" (OECD, 2008). A money mule works on behalf of the attacker and will receive and transfer the stolen money to the attackers, getting a commission in return (D'Alfonso, 2014). The money mules are important in the process because their presence will reduce the risk of the attackers from being exposed and identified (D'Alfonso, 2014). An interesting fact that Tajalizadehkhoob (2013) highlighted is that the money mules can be high school students who agreed to lend their debit card temporarily in exchange for a sum of money.

Security guardians

Security guardians are internal or external actors responsible for developing and enhancing the security of banking systems against the online fraud (Tajalizadehkhoob, 2013; Van Moorsel, 2016). The internal actors may include functions or specialised divisions under the bank organisation which are capable of or responsible for handling the bank's security, while the external actors are 3rd party security firms that partnered with the banks to provide their security services.

IT/network infrastructure service providers

This actor group refers to several types of service providers working around IT and network, especially Internet environment, for example Internet Service Providers (ISPs), hosting providers, cloud service providers, etc. Some of these service providers are potentially involved in the malware attack case and have capabilities to influence the situation. For example, hosting providers are capable to routinely take down C&C servers of the malware (Tajalizadehkhoob, Gañán, et al., 2017).

Regulatory body

The development of Internet banking and the prevalence of frauds in the Internet environment increase regulatory concerns related to it (Ezeoha, 2006). A regulatory body defined in this thesis is essentially an entity who has the authority to regulate the security of cyberspace especially if it links to the online banking process. It can be local, national, regional or global authorities. Due to the borderless nature of cyber threats, a collaboration between governments, regulators and industry and a high degree of alignment across national regulatory aspects is required (Crisanto & Prenio, 2017).

2.5. Criminology Theory for Analysing Factors Influencing Target Selection of Financial Malware Attacks

2.5.1. From Routine Activity Theory to Situational Activity Theory

The malware attack is an example of criminal activities happening in the cyberspace. Like other types of criminal activities, it is also studied in the field of Criminology. In order to understand how criminal activities can occur, criminologists have been developing and evaluating multiple theories and conceptual frameworks for decades. Amongst the many theories, some concepts stand out and are widely used to analyse criminal activities, namely Routine Activity Theory, Rational Choice Theory, Opportunity Theory (RAT + RCT combined) and Situational Activity Theory.

The RAT theory, developed by Cohen & Felson (1979), stated that there are 3 (three) minimal elements for conducting a successful violation: a motivated offender, a suitability target for the offender and absence of guardians. They also argued that the suitability of the target relates to the value, inertia, visibility and access (often abbreviated as VIVA) of the target. Value is related to traits of the target which offenders desire. Inertia refers to features of the target that could inhibit illegal actions of offenders towards it, for example the size of a property which disables a thief's ability to steal it. Visibility is related to the extent the target is known by offenders, while access refers to how easy the target could be reached. Figure 2.6 illustrates the RAT framework.

Meanwhile, Rational Choice Theory (RCT), proposed by Clarke & Felson (1993), has more em-

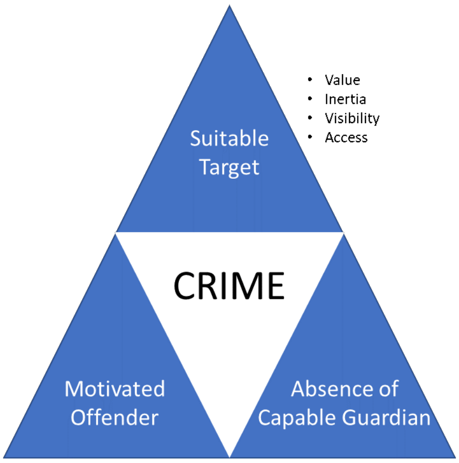


Figure 2.6: Routine Activity Theory by Cohen & Felson (1979)

phasis on the reason that drives people to do criminal activities. Its main propositions are that people's action decisions are purposeful, freely chosen and rational, and therefore drive people to choose an action aiming at optimizing outcomes in relation to their preferences.

Opportunity Theory Key tried to combine RAT and RCT by highlighting the role of RCT in explaining the motivated offenders part of RAT theory (Wikström & Treiber, 2016). It implies that RAT looks at the crime situation from a macro level while RCT looks at it from more of a personal level.

Finally, Situational Action Theory (SAT) emphasizes interaction between the person and environment, and relates these factors into the action mechanism (Wikström & Treiber, 2016). Its main highlight is the PEA (Propensity, Exposure, Action) hypothesis, which stated that an action comes out from a perception-choice process which results from the interaction between personal treats and stimulating external exposures. Figure 2.7 illustrates the mechanism of SAT.

2.5.2. Relevance of Routine Active Theory to Cybercriminals

Several prominent criminology concepts have been explained above and are potential to be used in the banking malware attack case. However, these concepts were mainly developed for the "physical" environment, raises a question of whether they can also be legitimately applied in the "virtual" environment to a certain extent (Leukfeldt & Yar, 2016).

There are two conflicting perspectives among criminologists regarding the resemblance between physical and virtual crime. The "transformationists" view the virtual environment as new, discontinuous with the terrestrial world. Capeller (Capeller, 2001) highlights that the new context offered by the cyberspace, like the dematerialization of body and anonimity, implies the need for a new criminological paradigm. Meanwhile, the "continuists" suggest that both virtual and physical criminality are fundamentally the same. As emphasized by Grabosky (2001), a great difference between "virtual criminality" and the physical counterpart is only the medium they use while the fundamental of the crime itself is still similar. Combining both views, there is a possibility to adopt the current "physical" criminology concepts to the virtual environment with some adjustments.

Routine Active Theory (RAT) is often mobilized to show that existing criminology resources can still explain cybercrime due to several factors (Leukfeldt & Yar, 2016):

- RAT theory has been widely utilised for analysing various forms of criminal behaviour.
- It has clear, analytical schema which enables it to be easily implemented to various scenarios.
- It provides clear hints for policy and crime-prevention.

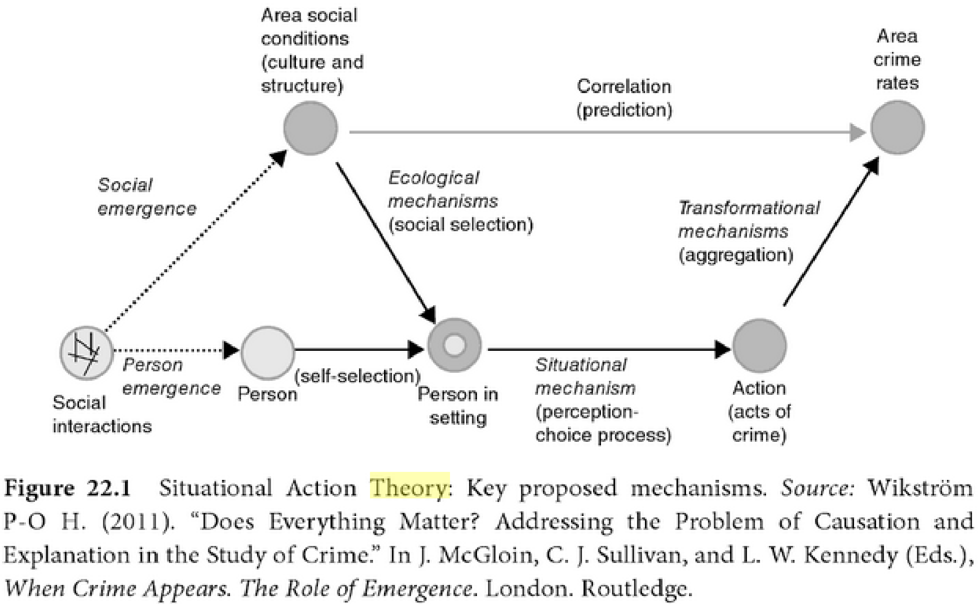
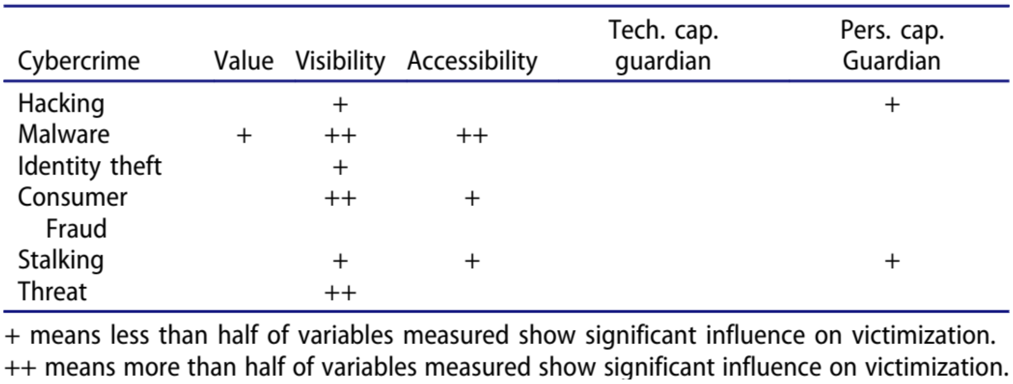

Area social conditions (culture and structure)
Correlation (prediction)
Area crime rates
Social emergence
Ecological mechanisms (social selection)
Transformational mechanisms (aggregation)
Person emergence
(self-selection)
Social interactions
Person
Person in setting
Situational mechanism (perception-choice process)
Action (acts of crime)

Figure 22.1 Situational Action Theory: Key proposed mechanisms. *Source:* Wikström P-O H. (2011). "Does Everything Matter? Addressing the Problem of Causation and Explanation in the Study of Crime." In J. McGloin, C. J. Sullivan, and L. W. Kennedy (Eds.), *When Crime Appears. The Role of Emergence*. London. Routledge.

Figure 2.7: Situational Action Theory (SAT) key proposed mechanism (Wikström, 2011)



Cybercrime	Value	Visibility	Accessibility	Tech. cap. guardian	Pers. cap. Guardian
Hacking		+			+
Malware	+	++	++		
Identity theft		+			
Consumer Fraud		++	+		
Stalking		+	+		+
Threat		++			

+ means less than half of variables measured show significant influence on victimization.
++ means more than half of variables measured show significant influence on victimization.

Figure 2.8: Effects of RAT parts in explaining cybercrime (Leukfeldt & Yar, 2016)

Based on a systematic reflection by Yar (2005), it is found that RAT is still capable to explain cybercrime as the core elements of RAT were shown to be applicable to the online environment. However, Yar (2005) argued that not all parts of RAT can be used in the cybercrime context. Inertia, for example, which refers to physical properties of matters that might generate resistance to a certain degree, is hard to transpose to the virtual environment. Yar (2005) pointed to files and technological specifications as an alternative to this. Leukfeldt & Yar (2016) also implied that not all parts of RAT can be used to explain cybercrimes. The representativeness depends heavily on the type of cybercrime analysed. Figure 2.8 shows the significance of parts of RAT in explaining the cybercrime. It shows that value, visibility and accessibility are three significant aspects for explaining the malware-related cybercrime. Based on these arguments, it is concluded that some aspects are still suitable for cybercrime, depending on the case. For the case as in this thesis, i.e. identifying the characteristics that influence the target selection of banking malware attack, three aspects of the RAT's suitable target element, value, visibility and accessibility, will be applied.

2.6. Identification of Characteristics Influencing Target Selection of the Malware Attack

This section describes the collection of several references which provide an explanation about the characteristics that may influence target selection of the malware attack. The references are collected by looking for previous theses and studies related to the target selection of banking malware, searching scientific articles from well-known scientific reference databases, such as Scopus and Google Scholar, with keywords like "target selection", "malware", "banking", "financial" and combination of them, as well as finding relevant technical documents from prominent security-related companies or institutions.

It was found that many of the currently available literature focuses mainly on the individual victims, which are the customers whose machine are infected by the banking malware, instead of the targeted banks. The reason is the targeted banks are not purely the direct victims in this attack case; their customers are the real victims. Yet, there exists several references describing the factors of the target selection from the banks' perspective. They are, therefore, used in this thesis.

From the collected references, the extraction of possible characteristics were performed. These characteristics were then linked to the relevant aspects of RAT's suitability target explained in the previous section (Value, Visibility and Accessibility).

Table 2.1 provides a structured list of the characteristics together with references that support it.

Table 2.1: Identification of Characteristics which potentially influence target selection

Characteristics	Evidence from the literature
Value	
Bank size, in terms of: • number of customers • number of online users • total assets • total payments • revenues • net profit	• (Tajalizadehkhoob, 2013), in terms of number of (online) users. She argued that the number of online users of a bank indicates the number of potential targets which can influence how criminals value the bank as a target. • According to Van Moorsel (2016), it can be expressed in many ways such as net profit, total assets, total payments, number of clients, etc. • An article from Beazley Breach (2016) described that hacking and malware attacks increasingly targeted smaller and more vulnerable financial institutions, particularly those with annual revenues of under $35 million. • On the other hand, an article from Kaspersky (2018) described that there is a potential sign that criminals nowadays are focusing a lot of their attention on targeted attacks against large companies.

Characteristics	Evidence from the literature
Country (where the bank is located), in terms of: • financial status • number of Internet users • rate of banking/shopping penetration • the degree of cooperation between financial institutions and law enforcement • the degree of cooperation between financial institutions • money transfer policies of the country • the number of banks in the country • the availability of money mules within the country	• (Tajalizadehkhoob, 2013) considered country’s financial status (like GDP), number of Internet users and rate of banking/shopping penetration are as an influence to the situation. • Van Moorsel (2016) argued that the context of a country playing a role is due to several factors like the degree of cooperation between financial institutions and law enforcement, the degree of cooperation between financial institutions, money transfer policies of the country, the number of financial institutions in the country, and the availability of money mules within the country. • Wueest’s (2016) research showed that there is variance in number of infections in a country. • An article from Kaspersky (2018) highlighted that more than half of all users attacked by banking malware in 2016 and 2017 were located in only ten countries.
Visibility	
Brand popularity	• Brand popularity is mentioned in an article by Kaspersky (2018). The argument is that top transnational banks, popular payment systems, Internet shops and auction sites become (cybercriminals’) favourite targets due to their brand popularity.
Domain name visibility	• Tajalizadehkhoob (2013) mentioned that "Banks with the word ’bank’ in their domain names" is argued to be a more likely target. However, the result of her analysis does not support this hypothesis.
Banks’ attack record	• This characteristic is mentioned by Tajalizadehkhoob (2013). According to her, a bank that has been successfully attacked has a higher tendency to be targeted in the future (as an effect of herding/information cascade). • Contrarily, Van Moorsel (2016) argued that the victims that were not under attack yet are more attractive.

Characteristics	**Evidence from the literature**
Website domain popularity	• Tajalizadehkhoob (2013) found that there is a weak and significant negative correlation between the domain popularity as indicated in Alexa ranking and the attack persistent to the domain. • Van Moorsel (2016) mentioned website rank in Google as a measure of website domain popularity which may influence the target selection.
Ownership of the bank	• Van Moorsel (2016) mentioned that the bank's ownership - public-owned or privately-owned - might influence its attractiveness as a target.
Language of the online banking	• Tajalizadehkhoob (2013) argued that banks offering English webpages might become more attractive. Her analysis showed that this hypothesis is valid only for the attacked entities within the EU region within the time period she focused on (< 2012). • (Cucu, 2017) argued that, although most financial malware initially targeted English users since those were primarily the wealthiest victims and also because of its international reach, there is an increasing trend showing localization likely to continue in the near future.
Accessibility	
Bank authentication method	• According to Tajalizadehkhoob (2013) interview with experts, a bank's authentication mechanism is believed to have a decisive role in determining access to a user's online account. • (Van Moorsel, 2016), as part of the "expected vulnerability" characteristic. • (Wueest, 2016), as there is an argument related to the authentication method applied. • (Kalige, Burkey, & Director, 2012), however, argued that even with the 2-factor authentication method, certain malware can still circumvent it, like in the case of Eurograbber they described.
Broadband penetration rate	• Both Tajalizadehkhoob (Tajalizadehkhoob et al., 2014) and Van Moorsel (2016) thought that broadband penetration rate of a nation might have a significant effect since the probability of people in a country going online increases as the Internet connection improves. Tajalizadehkhoob also concluded that this factor is significant at a worldwide level.
Users' online awareness	• Tajalizadehkhoob (2013) argued that the more knowledge online users have about the online environment and its threat, the less probable they will be of getting infected. • Van Moorsel (Van Moorsel, 2016) argued that the awareness of clients (about security in online transaction) reflected the vulnerability of financial institutions.

Characteristics	Evidence from the literature
Security control, in terms of: • Rate of use of firewall/antivirus products	• Tajalizadehkhoob (2013) argued that the use of firewall or antivirus products increases the degree of security measure for users, which may lower the probability of users getting infected. • Van Moorsel (2016) argued that the quality of firewall reflected the vulnerability of financial institutions and criminals tend to target the institutions which seems vulnerable.
Ease in securely performing criminal actions	• (Tajalizadehkhoob, 2013), in terms of the role of transfer policy and clearance time in reducing the speed and ease of money transfer to cyber criminals. • Van Moorsel (2016) highlighted the role of law enforcement in reducing the ease of transferring money to cyber criminals. • Wueest (Wueest, 2016) highlighted one of the global factors: "countries where international transactions are more difficult and may require local steps to launder the money" might become less of a target.

2.7. Selection of Characteristics for the Next Phase of Analysis

For the analysis, only some characteristics out of all the ones identified in the previous section were chosen. The selection was made due to some limitations which prevent the author from analysing them all. The criteria considered for making the selection are the novelty of the analysis, that is whether the characteristic has been analysed in previous studies, and the feasibility of collecting the data related to the characteristics. The feasibility of data collection takes into account the availability of data and the feasibility of collecting the data within a limitied period since the whole research should be completed within six months.

It is found that some of the bank's characteristics in the list have been assessed by previous researchers. Van Moorsel (2016) have tried to see whether an indicator which he used to approach the size of bank, the number of bank customers, could explain the target selection. He also assessed other characteristics like the country where banks resided, banks' attack record and the ownership of the bank. Tajalizadehkhoob (2013) has also looked at many characteristics like domain name visibility, the presence of English in banks' online banking, the country where banks resided (in which she looked at the country's GDP and infection rate), the popularity of the bank's website domain and the broadband quality. However, there are some factors which were highlighted by previous researchers but not yet analysed or proposed as future research. For example, Van Moorsel (2016) highlighted a hypothesis extracted from his expert interview that financial institutions with 2-factor authentication are targeted as much as ones with 1-factor authentication, which he did not assess due to his time and data constraints. It is interesting if a study can analyse some characteristics that were pointed out by previous research, but are not able to be assessed yet.

However, like previous research, time and data constraints still present. Some of the characteristics are hard to be assessed simply because the corresponding data difficult to be collected. For example, in order to describe the size of bank, one might need to find any of financial data like total assets, revenues or net profit. Unfortunately, not all banks disclose such data. For small, private-owned banks, the data become almost impossible to get. The data is also scattered around and not many services collected such data in one repository. Even if there is, it is expensive and still it cannot be guaranteed that the data is complete. Collecting information for the characteristics like rate of use of firewall/antivirus products are also difficult as banks might not be willing to share such information. Other characteristics might also hard to be analysed because they are not absolute factors. Brand popularity, for example, can be subjective and in the absence of any complete and plausible rating system, it is hard to judge which bank is more popular than others. It is also true for characteristics like users' online awareness and ease in securely performing criminal actions.

Therefore, in light of the considerations and constraints as explained above, the below characteristics have been selected for further assessment by this thesis:

- The language offered by the banks' online banking services. Despite the analysis performed by Tajalizadehkhoob (2013) regarding the presence of English, this thesis could broaden it by looking at other languages, especially those which are present within the EU area. The data can still be collected by observing the online banking entities manually, although it may require a long time.
- The authentication factor of banks' online banking services. As previous researchers have not delved into this characteristic yet, assessing the authentication factor potentially adds value to the target selection research.

Besides, some characteristics, like the country and the popularity of website domain, are also included in this research as control variables. Previous research highlighted the potential of these characteristics in explaining the target selection. Hence, they can serve as a control mechanism to reduce biases that may come from the analysis model containing the language and the authentication factor. In other words, they are included to avoid a rationale that any significance of the language or the authentication factor of an online banking in explaining the target selection is actually there because of an effect from the popularity of the particular online banking or the country the corresponding bank is located.

3

Data Collection and Preparation

This chapter will explain the collection and preparation of the main data for the thesis, list of banks and attack target data, as well as the external data which corresponds to the characteristics that want to be assessed.

3.1. Introduction of the Datasets

The first main dataset is the list of banks in the EU. The list is obtained from the list of financial institutions provided by the European Central Bank (ECB). This dataset is required to get the whole banking landscape in the EU. Some elimination and grouping were done in order to make the list relevant for analysis, which will be thoroughly explained in section 3.2.

The second main dataset is the attack target dataset. This attack target dataset came from the database provided by Fox IT for Delft University of Technology. Fox-IT is one of the world's leading security companies based in the Netherlands, mainly active in preventing and mitigating online threats as a result of cyber attacks, fraud and data breaches (Tajalizadehkhoob, 2013; Fox-IT, n.d.). The data came from malware configuration files, which Fox IT collected over time by creating a system (honeypot) that resembled a bot to emulate the malware. With such system, they can collect data of malware attacks in all around the world. The data are then stored in a database system (PostgreSQL) together with the analysis data of these entries (e.g. malware variants, time, entity they targeted, etc.) from separate tables.

The dataset provides information about the attack target URL as extracted from the collected malware configuration files, the bank name (presented in the dataset as the subentity name) identified from the attack target URL, its domain, its country, the variant of malware / malware family corresponding to the attack target and the time it is identified. With such information, it is possible to see which URL or bank or domain is targeted by the malware and when it is targeted. With further processing, it is also possible to find to what extent a certain URL/bank/domain is targeted. Therefore, this dataset is representative enough to get a sense of target selection of the banking malware on banks.

Even though the combination of the first and the second main dataset can already provide much insight about the target selection on banks in EU, there is still not any information that enables the analysis on the selected characteristics. Therefore, more datasets were collected for this purpose. These datasets are refered to as external datasets and their collection process will be explained later in this chapter.

3.2. Preparation of the List of Banks Dataset

3.2.1. Collection of the List of Banks Dataset

The scope of this thesis focuses on analysing banks in EU countries. Therefore, it is necessary in the beginning to obtain the list of banks registered in the EU. As explained in the introduction of this chapter, the legitimate list of registered banks in EU was obtained from the European Central Banks (ECB) open data (European Central Bank, 2018). Since the data is regularly updated, it is important to specify the version of the data used for this thesis. The data used in this thesis was obtained on March 21st, 2018.

3.2.2. Preparation of the List of Banks Dataset

The data lists all financial institutions in Europe which is under the ECB's area of control. There were initially 7,166 financial institutions on the list, divided into several categories such as central banks, credit institutions, money market funds and other institutions. This thesis focused on credit institutions as banks belong to this category. It reduced the number of entities to 6,234.

Furthermore, the data is filtered further to banks that have headquarters (HQs) within Europe as there are several foreign branches of banks on the list that have HQs outside Europe and therefore cannot be treated as European banks. The dataset contains information on the location of the HQ of banks which are foreign branches. From there, the banks whose HQ is outside Europe can be removed from the list. It resulted into 6,096 entities remained on the list.

3.2.3. Grouping of Similar Entities on the List and Processing the List Prior to Merging

In order to later identify which banks from the list were targeted and not targeted, the list of banks would be merged with the subentity names and domains obtained from the main dataset. However, the merging process was not as straightforward due to complications in the data:

- The subentity names mostly have different naming conventions to the formal bank names in the credit institution list. For example, the subentity name "Rabobank" apparently has the formal name "Coöperatieve Rabobank U.A.".

- Some subentity names do not refer to bank institutions but the IT providers which provide services to the banks. An example of this is the subentity "Fiducia and GAD IT" which mainly provides online banking solutions to Volksbank-Raiffeisen Bank in Germany.

- Some banks, which are initially perceived as one bank, are apparently different banks with a different website domain and a different Internet banking service. This is clearly apparent in the case of Volksbank in Germany. Volksbank does not belong to one bank but it consists of different banks which are not communicating with each other and have their own online banking services.

- On the contrary, there are multiple banks listed which can be grouped into one as they are branches of one bank and have only one online banking service for all. This case can be seen in Raiffeisen Bank Austria.

Additional approaches were performed before the merging process is performed, in order to handle the above complications. The approaches were:

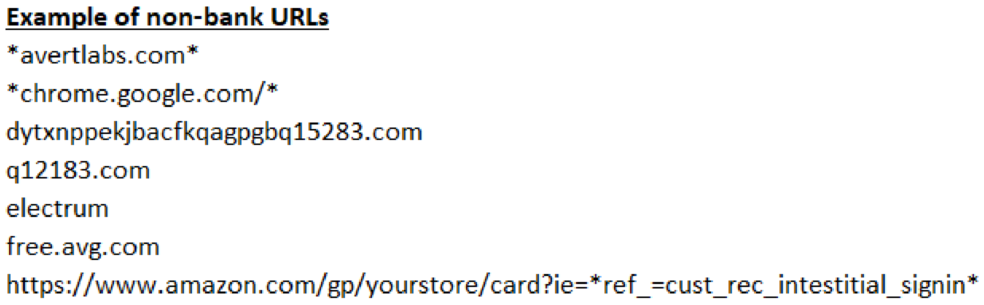
1. Banks whose headquarters are located outside EU were not considered in the analysis. This limitation is set because non-EU banks tend to have non-EU domains which are out of scope of this thesis. The banks' headquarters were clearly stated in the list of financial institutions from ECB.

2. Clustering was performed to group several banks in the same country, which were potentially branches of one bank, into one. The clustering process started with exploring the bank profile, for example collecting the information about its services, organisation, address and its e-banking domain. The grouping was performed afterwards, taken into account several criteria such as the bank name, its type (e.g. whether it is a consumer bank), its management and its e-banking domain. When it became certain that a cluster of banks was actually one bank and had a single e-banking service to accommodate the banks in that cluster, it was then considered as one single bank. This process was done in a mixed way, utilising both manual work and the clustering feature in Open Refine; an open source tool to refine and transform data (OpenRefine, 2018). An example of the grouping is presented in Table 3.1.

These approaches result in the list of 5,039 entities which will be merged with the attack target data.

Table 3.1: Examples of clustered banks

Bank	Country	Example of grouped banks	Reason of clustering
Raifeissen	Austria	Raiffeisenbank Region Baden eGen Raiffeisenbank Region Braunau eGen Raiffeisenbank Region Eferding eGen	Same domain for all branches (raiffeisen.at)
Caja Rural	Spain	Caja R. Central, S.C.C. Caja R. de Albal, C.C.V. Caja R. de Alginet, S.C.C.V.	Same domain for all branches (ruralvia.com)
Caisse d'épargne	France	Caisse d'épargne et de prévoyance Bretagne-Pays de Loire Caisse d'épargne et de prévoyance Côte d'Azur Caisse d'épargne et de prévoyance d'Alsace	Same domain for all branches (caisse-epargne.fr)
Credit Mutuel France	France	Caisse fédérale de crédit mutuel Caisse fédérale du crédit mutuel de Maine-Anjou et Basse-Normandie Caisse fédérale du crédit mutuel Océan	Same domain for all branches (credit-mutuel.fr)
BNP Paribas Spain	Spain	BNP Paribas España, S.A. BNP Paribas, S.E.	Same domain for all branches (bnpparibas.es)
Banque Populaire France	France	Banque populaire Aquitaine Centre Atlantique Banque populaire Grand Ouest Banque populaire Occitane	Same domain for all branches (banquepopulaire.fr)
Société Générale France	France	Société générale Société Générale SCF Société générale SFH	Same domain for all branches (societegenerale.fr)



Example of non-bank URLs

avertlabs.com

chrome.google.com/

dytxnppekjbacfkqagpgbq15283.com

q12183.com

electrum

free.avg.com

https://www.amazon.com/gp/yourstore/card?ie=*ref_=cust_rec_intestitial_signin*

Figure 3.1: Examples of non-bank URLs

3.3. Preparation of the Attack Target Dataset

3.3.1. Preparation of the Extracted Attack Target Dataset

As described in the introduction, the attack target dataset is originally stored in a database system. The dataset was extracted out of the database to make the next analysis easier. These steps were followed in order to extract the necessary data: analysing the structure and relationship of tables in the database, querying and extracting the query result into a comma-separated value (CSV) file. Due to confidentiality reasons, further details about these processes are not provided in this document.

The attack target dataset contains records from February 2014 to November 2017. Each record includes information about the URL targeted by the malware in an attack (attack target URL), the variant of the malware / threat that performed the attack, and the time. It represents a single attack target URL extracted from a configuration file of a variant of malware at a time. There are 14,198,778 records in the dataset, with 78,040 unique attack target URLs.

Some of the records also have information about the bank entity (domain, name and country) they targeted, based on their attack target URLs. The entity information is the result of an identification process done by Fox IT. Not all attack-target URLs in the dataset were URLs of banks; some of them belonged to other entities like social media platforms, IT services and antivirus companies. For these kinds of URLs, their entity information is not given simply because they are out of the scope of the identification mechanism. Some URLs can also be very random as they are in forms of regular expression and thus cannot be referred to any entity. Figure 3.1 provides examples of these non-bank URLs.

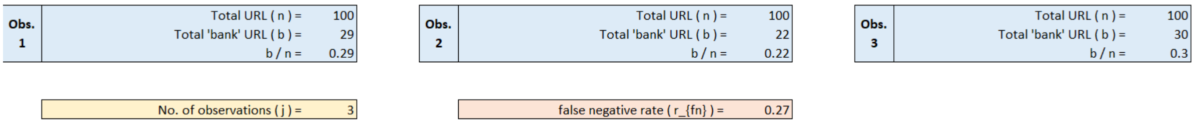
This thesis focused on analysing attack records which belong to banks within the scope of this thesis. The first approach of the analysis was to obtain the 6,800,729 records that have entity information from the dataset. These records can be further filtered so that only those belong to countries in the scope are taken into account. The countries are EU countries as listed in the bank list obtained from the European Central Bank (ECB). After filtering, there are 3,154,112 records, with 3,008 unique URLs, retained in the dataset for the next step.

3.3.2. Extraction of Entity Names and Domains

Besides the attack target data, the list of unique subentity profiles were also extracted from the database. Subentity domains, names and countries are required for the identification of targeted banks and for facilitating the merging of the dataset to the list of bank dataset. Merging is necessary as the full information of domains corresponding to each subentity is stored there.

3.3.3. Assessing the Mechanism of the Identification Process of Bank URLs

It is inferred from the dataset that Fox IT has a certain intelligence mechanism to identify bank entities from attack target URLs. In order to understand the mechanism better, an interview with a Fox IT personnel who took charge of the data was conducted.



Obs. 1		
	Total URL (n) =	100
	Total 'bank' URL (b) =	29
	b / n =	0.29

Obs. 2		
	Total URL (n) =	100
	Total 'bank' URL (b) =	22
	b / n =	0.22

Obs. 3		
	Total URL (n) =	100
	Total 'bank' URL (b) =	30
	b / n =	0.3

No. of observations (j) = 3

false negative rate (r_{fn}) = 0.27

Figure 3.2: Result of false negative test

It is inferred from the interview that Fox IT stored URLs of most of financial institutions, not only those of their customers but also of others that they explore themselves. In many cases, their customers provide all kinds of URLs they own, including the URL of their main page and their online banking. Some customers might only give the URL of their online banking. For those that Fox IT explore themselves, they only collect the online banking URLs. Meanwhile, it is important to notice that it was found later that some banks in the scope of this thesis do not have an online banking service. It will be explained later in section 5.5.

Each stored URL is also associated with the bank name and the bank's country of residence. Most of the time, their customers provide them with this information. If that is not the case, or for URLs they collected themselves, Fox IT uses a special approach in order to get such information. First, they use their knowledge or perform a research to find out about the bank associated with the URL and where it is located. Second, they check the top-level domain (TLD) of the URL and determine the location of the bank. For example, from the first step, the URL https://www.ingbank.pl/ will be associated with ING Bank which, according to their knowledge, resides in the Netherlands. However, from the second step, it is seen that its TLD refers to Poland. It is also clear from the website that it is ING Bank Poland. Therefore, they will associate the URL with ING Bank in Poland.

By matching the attack URLs (including those in regular expression format) with the stored URLs, they can identify the entities that belong to those attack URLs. There might also be the case that the attack URLs are so generic that they match nearly all entities.

The limitation in this mechanism is that most of the URLs came from Fox IT's customers and, unfortunately, not all banks in the EU are their customers. As a consequence, their mechanism might not able to identify all attack target URLs associated with all banks. This occurence is referred to as the presence of false negatives in this thesis.

In order to test the quality of this mechanism in terms of false negatives, a list of unique attack target URLs without entities was extracted from the dataset. There are 7,398,049 records without the entity profile (i.e. bank name and country), which comprise 72,902 unique URLs. False negatives were checked by randomly selecting 100 URLs from the list and checking how many of these URLs are actually bank URLs. This process was performed three times in order to ensure the fairness of the test. This test aimed to determine the false negative rate, which is defined by equation 3.1.

$$r_{fn} = \frac{\sum_{i=1}^{j} \frac{b_i}{n_i}}{j} \tag{3.1}$$

r_{fn} is the false negative rate, b_i is the number of URLs which apparently belong to banks in the i-th observation, n_i is total number of URLs in the i-th observation (100 for each observation) and j is total number of observation (3 times for this case).

Figure 3.2 illustrates the result of the performed false negative test. It is seen that the false negative for this dataset lies at the rate of 0.27, meaning that it is estimated that 27% of 72,902 unique URLs without entities are actually false negatives. This false negative rate is indeed high and can significantly influence the analysis in a way that introduces biases that prevent us from being able to conclude that banks which are not in the attack target data are not targeted.

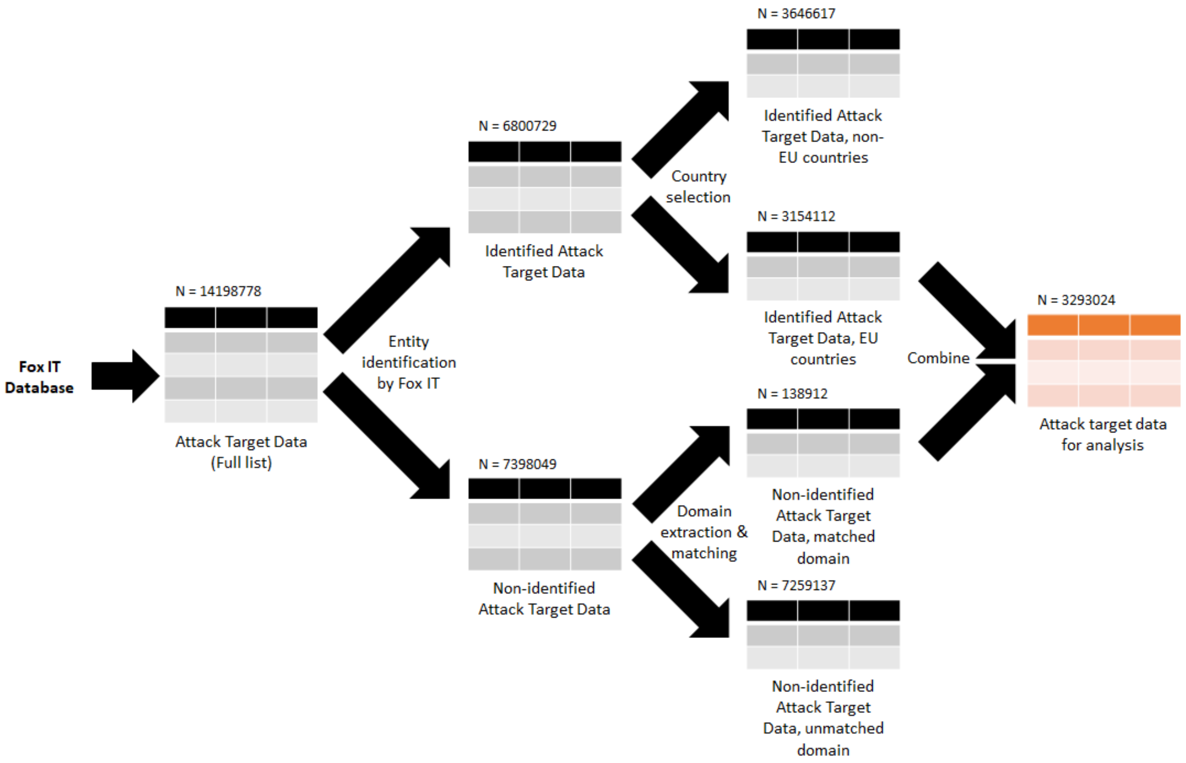

Fox IT Database
N = 14198778
Attack Target Data (Full list)
Entity identification by Fox IT
N = 6800729
Identified Attack Target Data
N = 7398049
Non-identified Attack Target Data
Country selection
N = 3646617
Identified Attack Target Data, non-EU countries
N = 3154112
Identified Attack Target Data, EU countries
Domain extraction & matching
N = 138912
Non-identified Attack Target Data, matched domain
N = 7259137
Non-identified Attack Target Data, unmatched domain
Combine
N = 3293024
Attack target data for analysis

Figure 3.3: Flow of attack target data processing until it is ready for analysis

3.3.4. Improving the Dataset by Domain Extraction

Due to the considerably high false negative rate in the attack target dataset, another approach was formulated to complement the first approach that was explained previously. The approach aimed to capture URLs in the false negative list which can be associated with any EU banks.

The approach was to extract domains from URLs that were not identified before. In order to do this, a procedure was coded in Python, utilising tldextract library, to obtain the second level domain of an attack target URL. To illustrate, if the attack target URL is https://www.xyz.nl/abc, the procedure will generate its second level domain xyz.nl. The code can be seen in appendix A.

The entity name and country of the extracted domains, however, were not available. In order to get the attack records that belong to banks in the scope of this thesis, the domains of the online banking services were collected manually for every bank in the ECB list. Attack records whose domains matched any domain in the list were then retained.

After the process, 138,912 additional records were found to be associated with any domain on the list, and hence were added to the dataset generated from the first approach, resulting in a total of 3,293,024 records being selected for the next phase.

While this approach was able to identify relevant records that were not identified by the first approach, there was still a limitation. This approach relied heavily on manual observation of domains that could be associated with the banks in the scope. There might be cases where not all domains of a bank can be collected as a bank can have multiple domains and some of them may be difficult to be searched. Human error can also affect this process.

Figure 3.3 illustrates the overall process flow of attack target dataset from the one extracted out of database to the one prepared for the analysis.

3.4. Merging the Attack Target Dataset and the List of Banks Dataset

As previously mentioned, the merging process cannot be done straightforwardly due to the complications in the datasets. Therefore, several processes in between were applied in order to ensure systematic and correct data merging. These in-between processes result in several temporary tables that aim to mediate both datasets. Each process is explained below.

Assign unique index to the list of banks

A unique number is assigned to each entity in the list of bank data. This unique number will serve as a key index in the merging process.

Assign unique index to the table of subentity names and domains

In the previous section, it is explained that the attack target dataset has information about subentity ID which indicates the entity that is targeted. There is also a table, extracted from the database, which maps subentity ID with corresponding entity name and domain. For each entry, a unique index is given which will serve as a key index in the merging process on the other side later on.

Create processed subentity-domain data table

This table is created due to complications in the previous table:

1. Multiple subentity ID may refer to the same subentity name (this indicates multiple domains)

2. Multiple banks may be under the same entity and need to be separated. This is true for Volksbank, Fiducia GAD & IT, and Credit Agricole

Due to these complications, the table of subentity names and domains is unable to be directly linked to the list of banks. To address the above complications, an intermediary table is included.

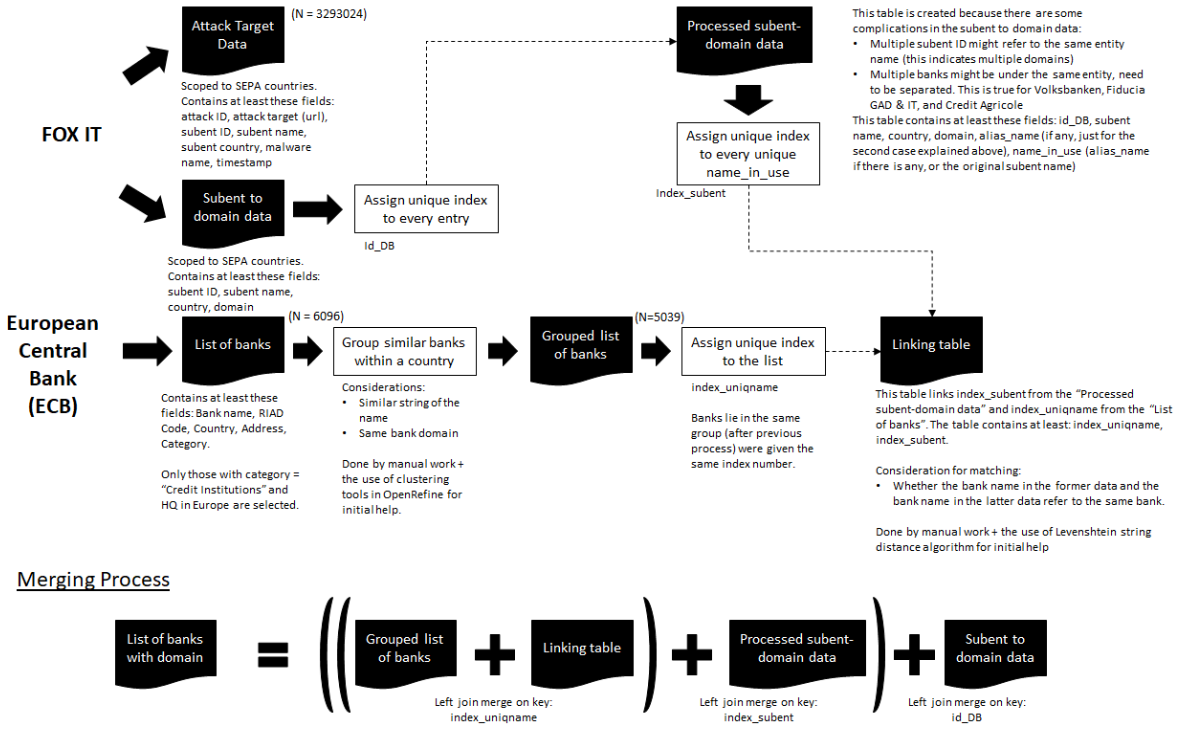
In order to better match the actual entities names stated in the banks list, i.e. the matching the dataset, several changes were made to the entities names, particularly for the above cases. Afterwards, a unique index is given for each entry for the key index in the merging process.

Create table linking the unique index of the processed subentity-domain data table and the unique index of the list of banks

This table serves its main purpose to actually link the list of banks on one side, and the attack target data (through its intermediary tables) on the other side.

In this process, the entity name in the attack target data and the entity name in the list of banks will be linked if they both refer to the same bank. The linking process is mainly done manually. The string distance algorithm "Levenshtein distance" was used to assist the mapping process. Levenshtein distance is a measure of the similarity between two strings where the distance refers to the number of deletions, insertions, or substitutions required to transform one string into another (Gilleland, 2006). The algorithm, unfortunately, cannot link all subentity and bank names correctly. Thus, manual check and correction were performed afterwards.

After entities on both sides have been linked to each other, the merging can be performed. Multiple merge processes were done, through the intermediary and linking tables, using the assigned unique indexes as the keys for the merging. Figure 3.4 illustrates the overall process of linking, creating intermediary tables, and the merging process of these two main data for the thesis in order to get the list of banks in EU with domains that can be linked to attack target data.


FOX IT
Attack Target Data
(N = 3293024)
Subent to domain data
Assign unique index to every entry
Id_DB
Processed subent-domain data
Assign unique index to every unique name_in_use
Index_subent
European Central Bank (ECB)
List of banks
(N = 6096)
Group similar banks within a country
Grouped list of banks
(N=5039)
Assign unique index to the list
index_uniqname
Linking table
Merging Process
List of banks with domain
Grouped list of banks
Linking table
Processed subent-domain data
Subent to domain data
Left join merge on key: index_uniqname
Left join merge on key: index_subent
Left join merge on key: id_DB

Scoped to SEPA countries. Contains at least these fields: attack ID, attack target (url), subent ID, subent name, subent country, malware name, timestamp

Scoped to SEPA countries. Contains at least these fields: subent ID, subent name, country, domain

This table is created because there are some complications in the subent to domain data:

- Multiple subent ID might refer to the same entity name (this indicates multiple domains)
- Multiple banks might be under the same entity, need to be separated. This is true for Volksbanken, Fiducia GAD & IT, and Credit Agricole

This table contains at least these fields: id_DB, subent name, country, domain, alias_name (if any, just for the second case explained above), name_in_use (alias_name if there is any, or the original subent name)

Contains at least these fields: Bank name, RIAD Code, Country, Address, Category.

Only those with category = “Credit Institutions” and HQ in Europe are selected.

Considerations:

- Similar string of the name
- Same bank domain

Done by manual work + the use of clustering tools in OpenRefine for initial help.

Banks lie in the same group (after previous process) were given the same index number.

This table links index_subent from the “Processed subent-domain data” and index_uniqname from the “List of banks”. The table contains at least: index_uniqname, index_subent.

Consideration for matching:

- Whether the bank name in the former data and the bank name in the latter data refer to the same bank.

Done by manual work + the use of Levenshtein string distance algorithm for initial help

Figure 3.4: Flowchart illustrating processes done to link and merge list of banks and attack target data

3.5. Collection of External Datasets

As described earlier in this chapter, some external data need to be gathered to enable the analysis on the selected characteristics. Therefore, the below data were collected:

1. The site language offered by banks’ online banking service. For example, whether its online banking offers service in English, its local language, or more than one language.

2. The authentication method of banks’ online banking service. For example, whether it only requires a username and password (1-factor authentication) or if there are additional features making it a multi-factor authentication, e.g. token.

3. The domain popularity ranking. This ranking can illustrate the relative popularity of a bank according to its online presence and is useful for controlling the model in the end.

3.5.1. Collection of the Site Language of Banks’ Online Banking Service

Information about the site language of banks’ online banking service was found by exploring the bank’s online banking service site. It was done manually as there was no secondary data showing this information and it is difficult to make an automation for it since there was no standard website design which enabled the collection of such information to be easily automated.

The checking was done by accessing the Internet banking website and looking for any language option offered there. For example, the ING Bank Netherlands’ website (www.ing.nl) was displayed in Dutch but offered the option to change the language into English. For that example, the website was perceived to offer two languages: Dutch and English.

3.5.2. Collection of Authentication Method of Banks' Online Banking Service Data

Similar to the site language, manual observation was done in order to justify the authentication method of the online banking service. Kiljan et al. (2018) described that there are two forms of authentication for the online banking: entity authentication and transaction authentication. The observation of transaction authentication, that is, user authorisation to a specific transaction in the online banking, is excluded in this thesis because of the following reasons:

- The European Banking Authority (EBA) has requested as per August 2015 that strong customer authentication is required for the customer's authorisation of Internet payment transactions, including transactions like money transfers using online banking services (European Banking Authority, 2014). Strong customer authentication is defined here as the use of 2- or multi-factor authentication. Therefore, the variability of transaction authentication between banks in Europe can be expected to be low.
- The observation of this type of authentication requires an internal access to every bank, for example by having an account there or by obtaining information directly from the bank. It is not feasible to perform this observation given the limited time and resources of this thesis.

Therefore, the observation was limited to the method for entity authentication (user login). The observation was done by considering several traits of the user login process:

- The display and structure of the login form itself. If it only contains username and password fields, it is more likely that it uses 1-factor authentication, which is "something that you know" (password). If the login form indicates the use of hard/soft token in addition to the password, then there is a high chance it uses 2-factor authentication ("something that you know" and "something that you have").
- The login help page and/or the Frequently Asked Questions (FAQ) page, if any. These pages often contain information or answers on the login procedure, which may be helpful in order to infer which authentication method an online banking service implements.
- The documents about the online banking services and/or the login instructions, if any. Similar to the point above, such documents may provide information about the authentication method used in a service.

A service might offer both 1-factor and 2-factor authentication for users to log into the system. For this case, both ways are recorded.

3.5.3. Collection of Domain Popularity Ranking

The ranking is obtained from Cisco Umbrella Popularity List, which reflects the relative Internet activity of a domain regardless of the invocation protocols and applications (Cisco Umbrella, n.d.). The list is updated every time, therefore the list from May 30th, 2018 is selected for this thesis. The list comprises the top 1 million popular domains in the world as per the time it is obtained. This list was then merged with the domains of online banking services so that it can be seen which domains are or are not in the list.

3.6. Concluding Remarks

This chapter described the collection and preparation of the data that will be used for this thesis. In general, five datasets were prepared: the attack target data, the list of EU banks, language (of an online banking service), authentication factor (of an online banking service) and domain popularity ranking. Summary of each dataset and the argument of its reliability and relevance to this thesis can be found in table 3.2. In short, these datasets are reliable enough and relevant for this thesis. The next chapter will describe more about metrics which can be extracted from these collected and prepared data.

Table 3.2: Summary of the collected data for this thesis

Dataset	**Source**	**Included Data**	**Reliability**	**Relevance to this Thesis**
Attack Target Data	Fox IT	Attack URL, entity profile (domain, name, country), malware variant, time	The data came from malware configuration files, which Fox IT collected over time by creating a system (honeypot) that resembled a bot to emulate the malware. The reliability of its entries can be guaranteed. However, its entity profile data is not fully reliable as false negatives are present.	Data in this dataset are needed for counting the attack.
List of banks in the EU	European Central Bank (ECB)	Bank name, country, office address, HQ profile	The dataset came from a legitimate entity. Therefore, it is reliable.	This dataset defines the population that wants to be researched in this thesis.
Language	Observation	Language offered by each online banking	Data were obtained from manual observation to every online banking website. Human errors which might have occurred during the observation can reduce the reliability of the data.	Language is one of the characteristics that want to be assessed in this thesis.
Authentication factor	Observation	Authentication factor implemented by each online banking	Data were obtained from manual observation to every online banking website. Human errors which might have occurred during the observation can reduce the reliability of the data.	Language is one of the characteristics that want to be assessed in this thesis.

Dataset	**Source**	**Included Data**	**Reliability**	**Relevance to this Thesis**
Domain popularity ranking	Cisco-Umbrella	Domain and its popularity ranking	The dataset was provided from a well-respected institution. The ranking was generated based on reliable measures, that is, based on combined measure of unique visitors and page views. Moreover, besides browser based ’http’ requests from users, this metric also considers the number of unique client IPs invoking a domain relative to the sum of all requests to all domains (Cisco Umbrella, n.d.). Thus, the ranking is an improvement from another similar ranking (i.e. Alexa) and its reliability can be guaranteed.	Domain popularity will be used as a controlling factor

4

Definition of Metrics for the Analysis

This chapter will address RQ2 by explaining the preliminary process of the explanatory analysis of this thesis: the definition of metrics that will be used in the analysis. The definition is important as it determines what will be measured and analysed.

Various metrics can be extracted from datasets that have been collected. These metrics correspond to the attack count, which can be used to indicate the target selection, and also to some characteristics that have been selected in section 2.7. These metrics will be useful for the analysis part, especially the explanatory analysis. This section also provides argument about the drawbacks of the existing approach of counting the attack and proposes some alternatives which are perceived to approach the actual attack count better.

4.1. Defining Attack Count

A metric that can be obtained from the attack target dataset is how many times a domain of an entity is attacked by a malware within the scope period. The basic definition of attack count is derived from this and this definition is often used by security companies like Symantec (Tajalizadehkhoob, 2013). However, Tajalizadehkhoob (2013) argued that this basic definition may not represent the valid number of times a domain is actually attacked. The next subsection will explain why this definition is not a good definition to count actual attacks. The next subsection will provide some argumentations about the basic definition and several alternative definitions proposed for this thesis in order to assess how these definitions may differently describe the target selection landscape.

4.1.1. The Drawback of the Basic Definition of Attack Count

The configuration file for a botnet can be updated by the bot master due to multiple reasons e.g. a mistake in the previous configuration, change in attack attempts, human errors, changes in antivirus software which force the bot to be updated to evade detections, etc. This update was captured as a new entry in the data although it did not actually correspond to a new attack. This may cause the overcounting problem in the existing count. Using the existing definition of attack count, an attack with two updates is going to be counted as three different attacks, while it is supposed to be counted as one. Take for example the illustration as in figure 4.1. It is seen that attacks to mijn.ing.nl are captured from three different configuration entries. However, the attacks in the second and third configuration entries are actually the same as in the first configuration entry. This is due to the second and third configuration entries serving as updates to the first configuration entry, resulting in the attack to mijn.ing.nl being left unchanged.

By using the current definition (raw count), attacks to mijn.ing.nl will be counted as 3, one from each configuration entry. However, it supposed to be counted as one attack only. This is the reason why overcounting occurs when the metric raw attack count is used.

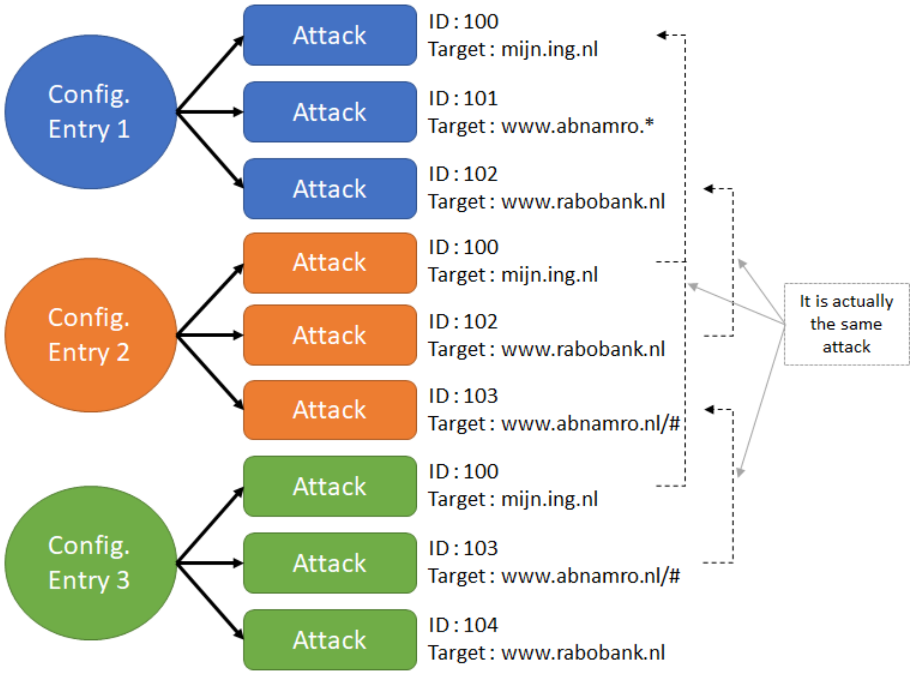


Figure 4.1: An illustration to show how raw counting may inflate the actual attack count

It is indeed not easy to determine the actual attack out of the list of entries as there are no information available in the data about which command-and-control (C&C) server sent the configuration, which botnet received it or which group of criminals made it. An attempt using time dimension as the key was made to approach the actual attack count, by assuming that criminals need some time to make a new attack attempt (Tajalizadehkhoob, 2013). However, a specialist in Fox IT questioned this assumption, arguing that it really depends on the malware family. If a malware family is used by many actors - like in the case of ZeuS malware, which is accessible to the public - it is possible to see the same target being attacked by the same type of malware at the same time (or with a very little difference in time).

According to the specialist in Fox IT, an attack has its own inject code and will not change despite an update of the configuration file. Therefore, he argued that different inject codes indicate different attack attempts. Fox IT has assigned a unique attack ID to each unique attack's inject code they collected. Two entries will be given different IDs if their inject codes are slightly different, even if they have the same attack target URL. As illustrated by the specialist, this case is like a thief trying to rob the same house using different means.

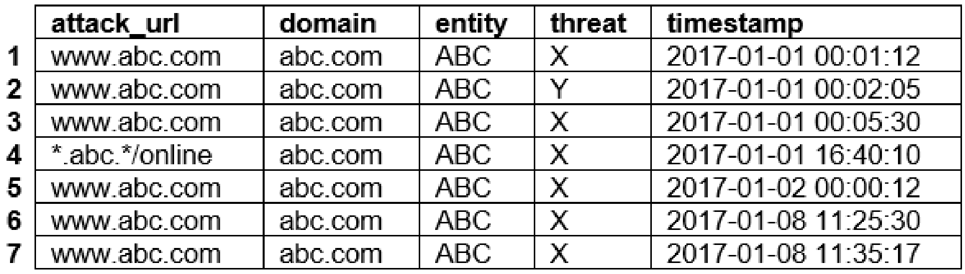
The next subsections will describe possible metrics that can be extracted from attack target data. Some of these metrics may potentially explain the actual attack better than the raw count. In addition, raw attack count is still included as a metric that can be extracted from the data in order to see and compare its performance.

4.2. Possible Metrics Extracted from Attack Target Data

4.2.1. Raw Attack Count

This thesis discusses argumentation about the basic definition vs alternative definitions to allow for them to be evaluated against each other. A raw attack count in this context refers to the number of attack URLs that refer to a domain of an entity over a certain period.

In order to illustrate how this definition works, a dummy attack target data which resembles the



	attack_url	domain	entity	threat	timestamp
1	www.abc.com	abc.com	ABC	X	2017-01-01 00:01:12
2	www.abc.com	abc.com	ABC	Y	2017-01-01 00:02:05
3	www.abc.com	abc.com	ABC	X	2017-01-01 00:05:30
4	*.abc.*/online	abc.com	ABC	X	2017-01-01 16:40:10
5	www.abc.com	abc.com	ABC	X	2017-01-02 00:00:12
6	www.abc.com	abc.com	ABC	X	2017-01-08 11:25:30
7	www.abc.com	abc.com	ABC	X	2017-01-08 11:35:17

Figure 4.2: A dummy attack target data

actual dataset used for this thesis is shown in Figure 4.2. This dummy table presents 7 attack entries which are assumed to be all attack entries referring to the domain of the entity ABC in the first two weeks of January 2017. For indexing purpose, numbers are assigned to each of the entries.

According to the explanation above, the basic definition will certainly consider all of the attack entries in the dummy data as actual attacks, thus the attack count for this example will be 7. However, it is highly unlikely that all of those entries are actual attacks. For example, the third and fifth attack entries are present not long after the first attack entry, suggesting that they may be updates for the first attack entry.

Applying the definition to the attack target dataset

The definition was applied on the attack target dataset to see how it performs. The observation focused on finding the top 10 most targeted banks and the cumulative distribution of percentage of attack counts over percentage of total banks in the data, according to this definition.

Figure 4.3 lists the top 10 most targeted domains according to raw counts. Based on the definition, it is seen that Volksbank Raiffeisenbank Nordoberpfalz in Czechia is the most targeted, followed by Deutsche bank in Germany, while the first outnumbered the second-ranked by almost twice. However, this attack count does not necessarily imply the count of actual attacks, due to the limitation explained previously. Meanwhile, also from the same figure, it is seen that, similar to a finding by Tajalizadehkhoob (Tajalizadehkhoob, 2013), the pareto power law still applies: 20 percent of total banks in EU accounts for more than 80 percent of total attack count.

Advantage

The raw count is the simplest approach to count the attack while still able to provide insights about the target selection in a rough case (e.g. the top targeted domains).

Limitation

Some entries are not meant to make a new attack to the particular domain and instead may only be updates to the previous version. Therefore, there is a chance that the count is actually inflated.

4.2.2. Number of "Week-interval" Attack Count

This definition is based on a rationale that it is unlikely to perform new attacks that have very similar configuration file characteristics (same attack target URLs using the same malware variant) at a very short time and, if there are such entries, the most rational argumentation is that they are meant to update the previous configuration file. In line with the argument that the bot master may update the configuration once every two days at the most frequent (Tajalizadehkhoob, 2013), this definition assumes that new attacks that have very similar characteristics as those from the past may occur only 7 days (a week) after the previous similar attacks. Hence, this definition tries to eliminate the

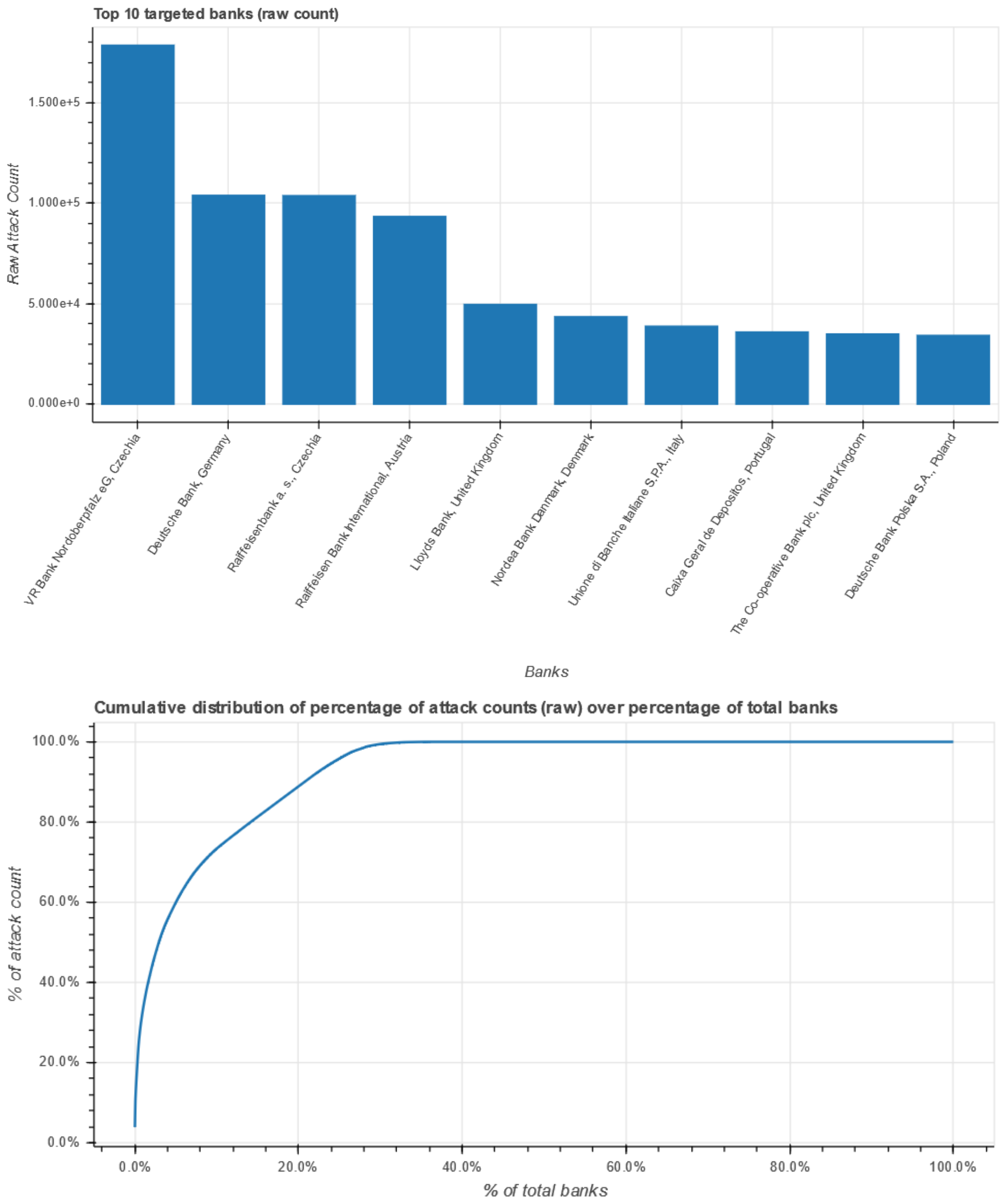


Figure 4.3: Top 10 most targeted banks (top) and cummulative distribution of percentage of attack counts over percentage of total banks (bottom), according to the raw attack count

entries that may be associated with updates to the previous attack configuration.

Looking back at the dummy data in Figure 4.2, this definition will eliminate entry 3, 5 and 8. Entry 3 and 5 happened in less than 7 days, even less than 24 hours, after the first entry, which shares the same traits. The same reason applies to entry 7 against entry 6. Entry 6, however, is not eliminated since it happened 7 days after the perceived actual attack of entry 1. Therefore, this definition leaves 4 entries that are perceived to be actual attacks. This reduces the chance of overcounting the attack as the entries which seem to update the actual attack will be eliminated and not counted.

Applying the definition to the attack target dataset

The week-interval definition was applied to the attack target dataset in the same manner as the previous definitions. Figure 4.4 lists the top 10 most targeted domains and also the cumulative distribution, according to this definition.

It is seen that, compared to the raw count, there is a different order of the top 10 targeted banks. Raiffeisenbank a.s. in Czechia became the most targeted banks while Raiffeisen Bank International in Austria was the second. Deutsche Bank Germany, while becoming the second most targeted by the raw attack count, became the fourth by this definition. The cumulative distribution still maintain a similar pattern, while 20 percent of total banks accounts for more than 80 percent attack count.

Advantage

Suppose the assumption about attack interval is right, it can provide a significantly more accurate count of the attack targeted to a domain since it will ignore any entry that occurred to update the previous attack.

Limitation

The metric relies on the assumption that the next actual attack will not occur before a week. If there is a new attack with the same entry characteristics as the previous one which happen for less than a week, it will not be taken into the count.

In addition, it is argued before that the malware family heavily influences the assumption about the time interval. The current assumption, 7 days, was extracted from a study which uses ZeuS data from a period before 2013. This assumption may not be relevant anymore to the current data and to other malware families. In order to make certain that the attack entry is the same as the previous entry, one should check if the corresponding C&C server is sending the related configuration file to the same bot at the same time. If not, this assumption cannot be applied.

4.2.3. Count of Unique Attack ID

This metric is created based on the previous argument that an attack ID which is present in the dataset refers to a unique inject code and hence may indicate different attack attempts performed by criminals. This metric counts the number of unique attack ID which corresponds to a domain. This count may approach the actual attack count better than raw count because it only considers multiple attack entries which have the same inject code - represented by the same attack ID - as one entry. As previously argued, multiple entries with the same inject code may happen due to updates of configuration files and they would be the cause of the overcounting problem in the raw attack count metric.

Applying the definition to the attack target dataset

This definition was applied to the attack target dataset in the same manner as the previous definitions. Figure 4.5 lists the top 10 most targeted domains and also the cummulative distribution, according to this definition.

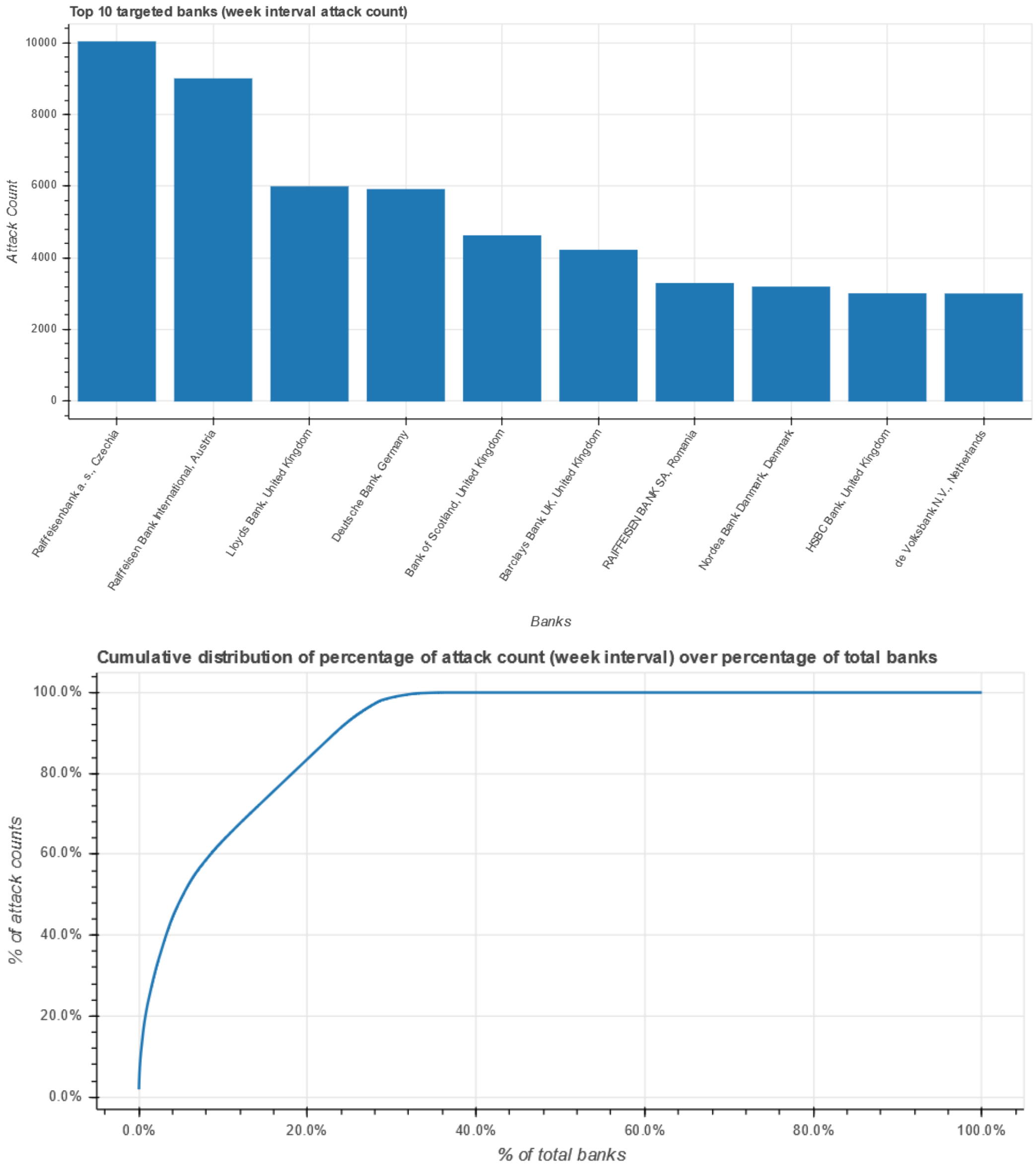


Figure 4.4: Top 10 most targeted banks (top) and cummulative distribution of percentage of average number of attack over percentage of total banks (bottom), according to week-interval attack count metric

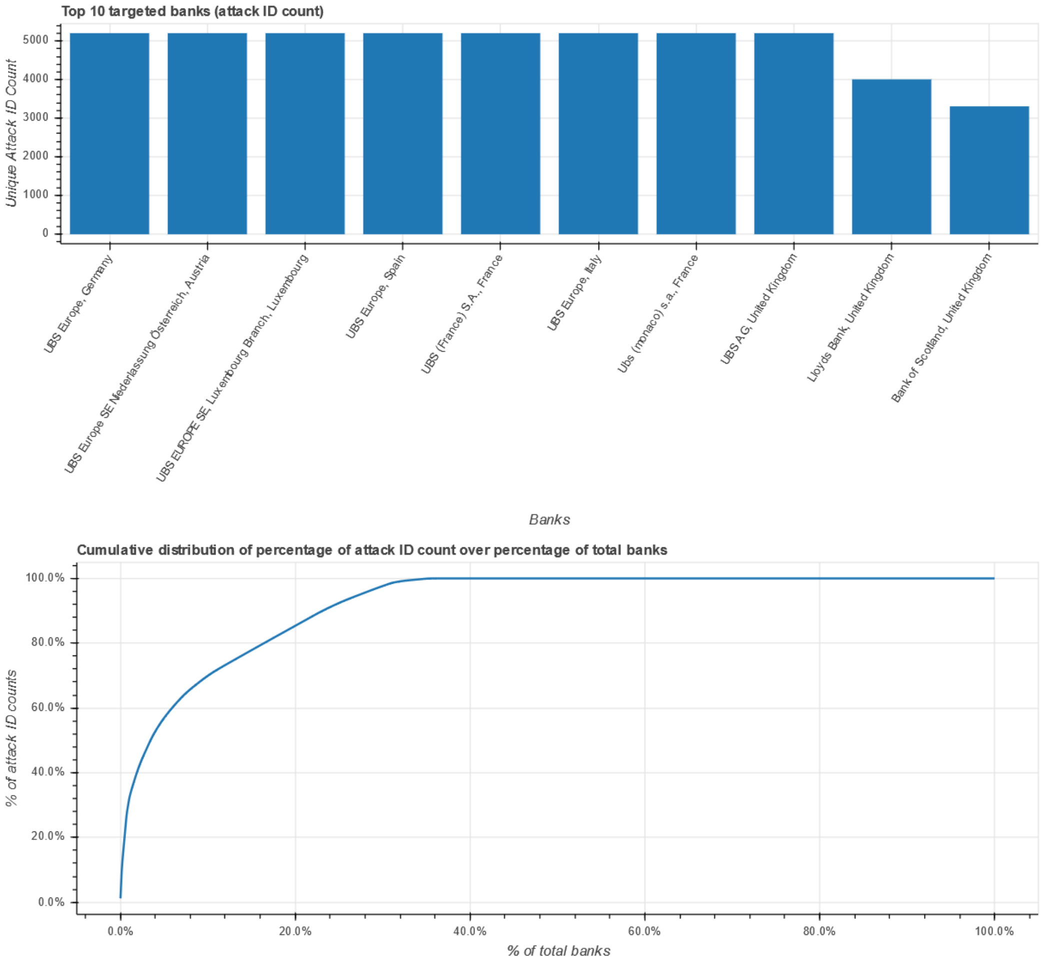


Figure 4.5: Top 10 most targeted banks (top) and cummulative distribution of percentage of average number of attack over percentage of total banks (bottom), according to unique attack ID count metric

It is seen that the top 10 targeted banks according to this metric is different than the one from previous metrics. Banks under UBS Group appear as among the most targeted banks, followed by banks in the United Kingdom. The cumulative distribution still maintain a similar pattern: 20 percent of total banks accounts for more than 80 percent of attack count.

Advantage

Should the argument that different inject codes indicate different attack attempts be true, this metric can significantly eliminate the overcounting problem that the raw attack count metric encounters. Therefore, one could be sure that the count from this metric represents the actual number of attacks targeting a bank.

Disadvantage

Different attack attempts that share the same inject code may still be present although the chance is very small.

4.2.4. Number of Weeks an Online Banking Service is Under Threat

A metric that can be extracted from the attack target data and could potentially describe the rate of target is the number of weeks a bank is under threat. This metric can in general describe the persistence of an attack towards an entity.

This metric can be defined in two ways (Tajalizadehkhoob, 2013): (1) the number of weeks when any attack entry is present (meaning the configuration file received by a botnet) and (2) the number of weeks as in the first definition as well as the gap weeks between two attack entries, following the assumption that the botnet was also active in these intermediate weeks. While Tajalizadehkhoob (2013) argued that the second assumption is closer to reality, the second definition leaves an uncertainty about the gap weeks since one cannot ensure if the attack is actually stable during the gap weeks or if there is no longer an attack to the entity. Moreover, it is not rational to assume that the attack still occurs when the gap weeks are very long. This makes the reliability of the second assumption questionable. Therefore, in contrary to Tajalizadehkhoob, this thesis will follow the first definition.

There are 192 weeks in total in the period observed by this thesis (February 2014 - November 2017). Some banks have been found to be targeted persistently in a whole period. Referring to the top 10 banks with the longest weeks under threat in the period (table 4.1), many of the persistently-targeted banks are in the United Kingdom. However, figure 4.6 suggested that the persistence of the attack is different for every bank and is quite evenly distributed. Many banks have 0 weeks under attack, as many banks are identified to never be targeted.

Advantage

This metric can show the attack persistence to an online banking entity. One can say that the more persistent is the attack, the more they are perceived to be targeted by the criminals.

Disadvantage

This metric cannot provide any insight about the intensity of the attack received by the online banking entity. An online banking receiving a thousand attacks distributed in two weeks will be considered less targeted than an online banking receiving only ten attacks distributed in two months.

4.2.5. Number of Unique URLs Corresponding to the Bank

The number of unique URLs can also become an indicator of a bank's target rate. The rationale is it is unlikely for two completely different attackers to use exactly the same URL to target an online banking entity. If they actually want to target an online banking entity, they will target at least a

Table 4.1: Top 10 banks with the longest weeks under threat

Bank	Country	No. weeks under threat
The Royal Bank of Scotland	United Kingdom	195
The Co-operative Bank plc	United Kingdom	195
TSB Bank Plc	United Kingdom	195
Lloyds Bank	United Kingdom	195
Lloyds Bank	Netherlands	195
HSBC Bank	United Kingdom	195
Bank of Scotland	United Kingdom	195
National Westminster Bank	United Kingdom	195
Barclays Bank	Italy	194
Barclays Bank	United Kingdom	194

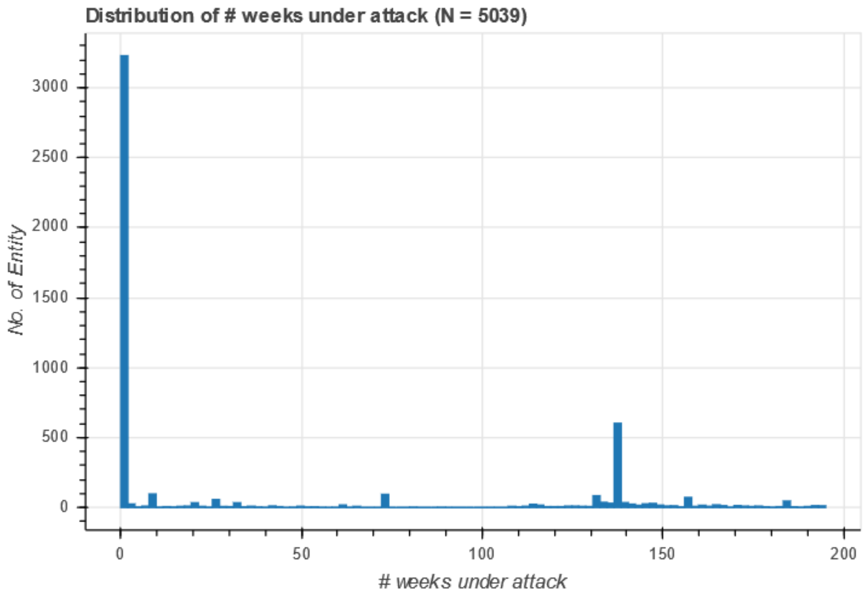


Figure 4.6: Histogram on number of weeks banks are under threat

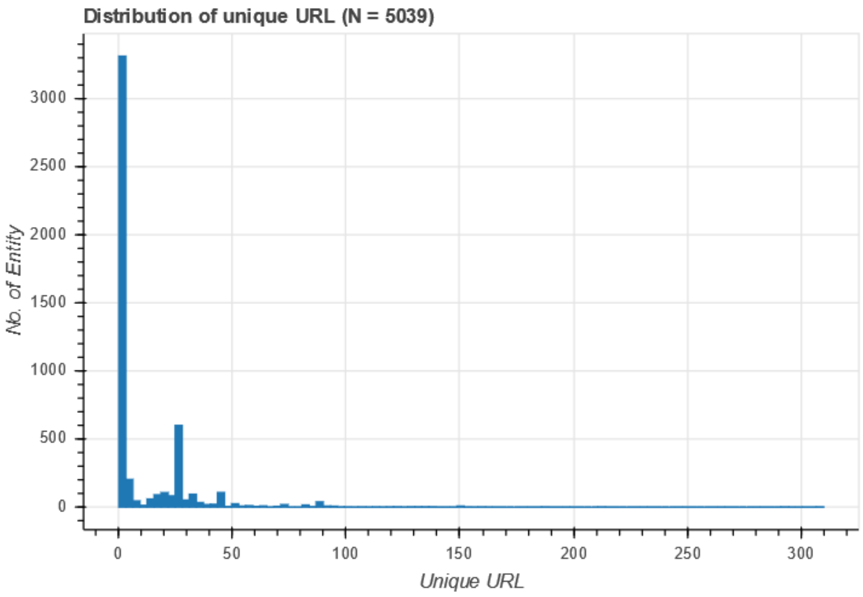


Figure 4.7: Histogram on number of unique URLs associated with an online banking

Table 4.2: Top 5 banks with the most unique URLs corresponding to them

Bank	Country	No. unique URL
Lloyds Bank	United Kingdom	310
Bank of Scotland	United Kingdom	294
Barclays Bank UK	United Kingdom	211
HSBC Bank	United Kingdom	195
BANCO DI NAPOLI S.P.A.	Italy	161

different part of the online banking entity or use different ways to express the attack destination (for example, using regular expression), resulting in different URLs. Therefore, more unique URLs associated with an online banking could mean higher actual target rate for the online banking.

Table 4.2 presents the top 5 banks with the most unique URL associated with them. It is seen that many United Kingdom banks are on the list. Like the metric of the number of weeks an online banking service is under threat, the number of URLs are quite evenly distributed seen in figure 4.7. Most banks have less than 50 unique URLs associated with them.

4.2.6. Number of Different Malware Cariants Targeting an Online Banking Service

As the attack target dataset contains information about the threat or the malware variant that corresponds to an attack entry, a metric involving the threat information can be extracted from the dataset.

There are 29 unique threats captured in the dataset. Each threat is used differently, targeting different entities in different countries. Some are used persistently over time while some only for specific attacks. It is also possible for multiple threats to target the same online banking entity during the period. Like unique URLs, more threats targeting an online banking entity could suggest that the online banking entity is more attractive for attackers since it implies that different malware users were looking at the same online banking as the prospective target of their attacks.

Table 4.3: Top 5 banks with the highest number of threat targeting them

Bank	Country	No. unique URL
Lloyds Bank	United Kingdom	23
Bank of Scotland	United Kingdom	23
HSBC Bank	United Kingdom	23
ING Bank	Netherlands	22
TSB Bank Plc	United Kingdom	21

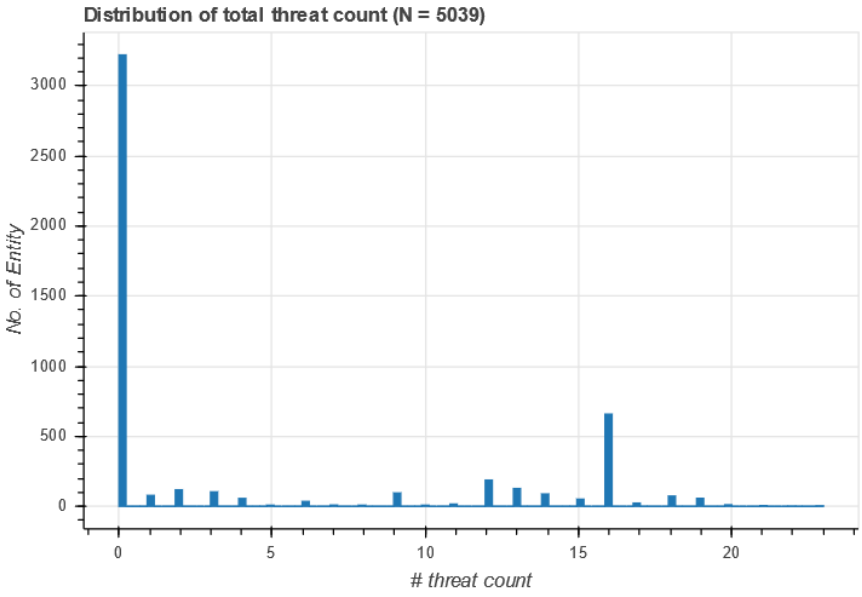


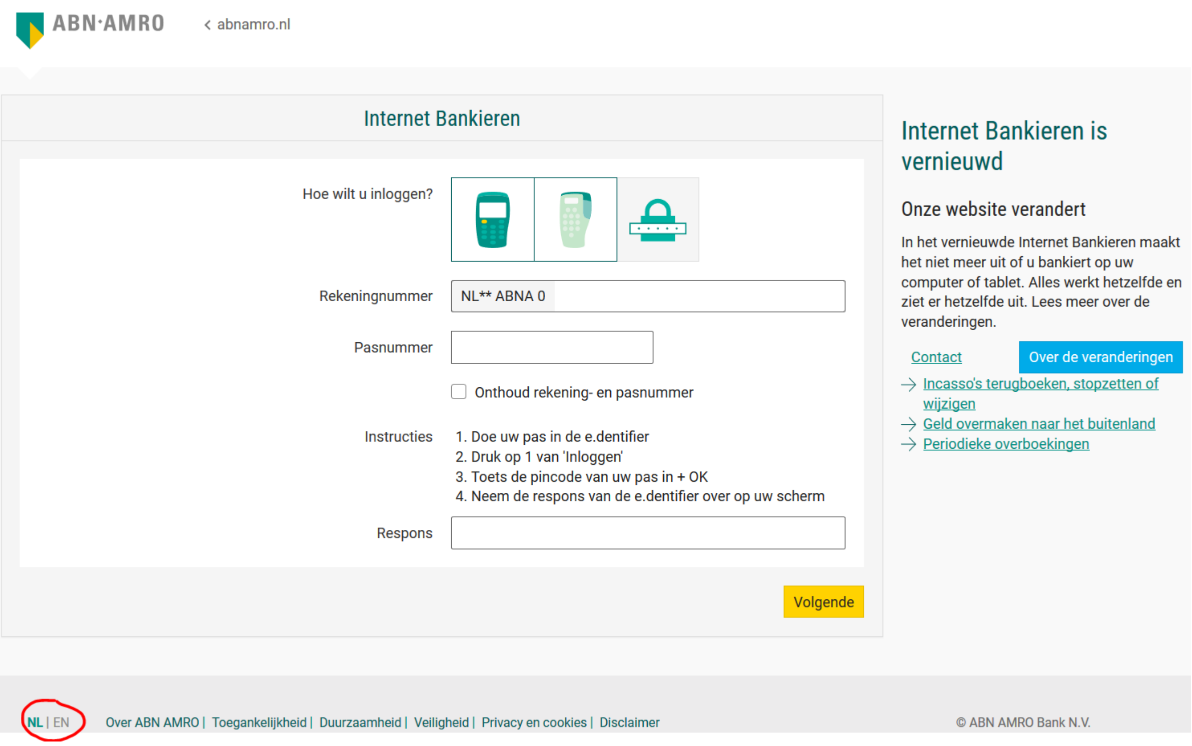
Figure 4.8: Histogram on number unique threats targeting an online banking

Table 4.3 presents the top 5 banks with the highest number of threats targeting them. Like the unique URLs metric, many United Kingdom banks are on the list. It is seen that there are no banks in EU that are targeted by all malware variants available in the dataset. Like the metric of number of weeks an online banking service is under threat, the number of malware variants is quite evenly distributed, as can be seen in figure 4.8. Amongst the targeted banks, many were targeted by less than 17 different threats.

4.3. Possible Metrics Extracted from External Data

4.3.1. The Scope of Online Banking Service

It was illustrated in the beginning of Figure 2.2 that online banking services nowadays may be offered in multiple forms, such as website (home banking), mobile site and mobile app. Subsection 2.2.2 has also explained why assessing cyberattacks on the home banking websites is still relevant compared to other channels. Therefore, this thesis will focus on the home banking website as its scope for online banking service and only considers the external parameters that correspond to banks' home banking services.



ABN·AMRO

< abnamro.nl

Internet Bankieren

Hoe wilt u inloggen?

Rekeningnummer NL** ABNA 0

Pasnummer

□ Onthoud rekening- en pasnummer

Instructies
1. Doe uw pas in de e.dentifier
2. Druk op 1 van 'Inloggen'
3. Toets de pincode van uw pas in + OK
4. Neem de respons van de e.dentifier over op uw scherm

Respons

Volgende

Internet Bankieren is vernieuwd

Onze website verandert

In het vernieuwde Internet Bankieren maakt het niet meer uit of u bankiert op uw computer of tablet. Alles werkt hetzelfde en ziet er hetzelfde uit. Lees meer over de veranderingen.

Contact Over de veranderingen

→ Incasso's terugboeken, stopzetten of wijzigen
→ Geld overmaken naar het buitenland
→ Periodieke overboekingen

NL | EN Over ABN AMRO | Toegankelijkheid | Duurzaamheid | Veiligheid | Privacy en cookies | Disclaimer © ABN AMRO Bank N.V.

Figure 4.9: ABN Amro N.V.'s online banking website. The red circle highlights the ability of the visitors to change the language of instruction of this website.

4.3.2. Metrics from Language Data

One of the external data collected for analysis is the language of the online banking service. The language in this context is defined as the human language(s) which is/are offered by the online banking service site.

The language of an online banking service site can be easily identified by only looking at what language is displayed when it is opened. For example, the online banking service site of ABN Amro N.V., one of the largest banks in the Netherlands, is seen to offer Dutch language in the first instance, as can be seen in Figure 4.9. Some online banking sites, however, may offer more than one language. It can also be seen in the case of ABN Amro in which the customer can alter the language of instruction from Dutch to English by clicking at the link at the left-bottom side.

From this language data, some metrics can be extracted. They are described in the following points.

The presence of a language in an online banking entity

The metric is straightforward: if a language is present in an online banking service, the metric will have a True value, otherwise False. The metric can be created for each language that is available in the data. Eventually, from such metric, the relationship between the presence of a certain language and the target rate can be assessed.

Number of languages offered by an online banking entity

Many online banking sites outside english-speaking countries offer at least two language options: English and their local language. Some may offer more due to the proximity of their locations. Others, however, may only offer one option. This variety is also interesting to look into as it might also influence the target selection of banking malware.

4.3.3. Metrics from Authentication Factor Data

The authentication factor in this context refers to that used to prove the identity of an online banking user (entity authentication). Revisiting the theory of authentication method, there are at least three possible factors for the authentication, such as something the user knows (knowledge), something the user has (possession) and something the user is (biometrics) (Claessens et al., 2002; S. Z. Kiljan, 2017; Tajalizadehkhoob, 2013). A method is considered one-factor authentication (1FA) if it utilizes only one of these factors. The most common 1FA is the simple combination of static username and password. On the other hand, a method is considered two-factor authentication (2FA) if it combines two of these factors.

In the example of ABN Amro above, it is seen from its login page that it uses a physical device called "e.dentifier" combined with other necessary details in order to enable the user to log into the system. Therefore, its authentication method is considered as a two-factor authentication (2FA). However, further exploration found that it also allows the user to log in with a 5-digit passcode and without the e.dentifier. Therefore, besides 2FA, ABN Amro still maintains one-factor authentication (1FA) method too. It is not uncommon to see banks offering several methods of user authentication.

A metric that can be extracted from the data is **the presence of a certain authentication factor in an online banking site**. Similar to the metric for language, if an authentication is present in an online banking service, the metric will have a True value, otherwise False. With such metric, the relationship between the presence of a particular authentication factor and the target rate can be assessed.

4.3.4. Metrics from the Domain Popularity List

The domain popularity ranking from Cisco Umbrella includes the one million most popular domain worldwide. A simple metric can be generated from this ranking, which is **the ranking of an online banking domain**. Simply put, the metric will contain the ranking of the online banking domain as on the list.

4.4. Concluding Remarks

This chapter identified and described metrics that can be extracted from the collected data. Several approaches to count the attack, which are important in order to quantify the rate of target selection to an online banking entity, were provided. Each approach has its own advantages and disadvantages, therefore, it is difficult to justify which approach can lead to the perfect counting of actual attacks. However, this thesis argues that the attack ID count is better than other similar metrics in approaching the actual attacks as the metric is based on the confirmation of Fox IT that a unique attack ID relates to a unique inject code, which indicates a unique attack. For the next analysis, especially in the explanatory analysis, these three metrics are maintained: raw attack count, week-interval attack count and attack ID count, so that the performance of each metrics can be seen and evaluated.

Besides the attack count metrics, other metrics were also extracted from both main and external datasets to be used as independent variables in the explanatory analysis.

5 Descriptive Analysis

This chapter will describe the data which were already collected and provide some interesting insights about the target landscape of online banks that can be extracted from these datasets, answering RQ3 of this thesis. There are many findings that can be gathered regarding the target selection of banking malware from the collected data, with some data transformation and visualisation.

5.1. Basic Description of the Data

As mentioned in the previous chapter, after filtering and grouping the list of banks obtained from the European Central Bank (ECB), 5,039 banks throughout European Union (EU) remain for the analysis. The banks are spread across 28 EU countries[1]. Figure 5.1 presents the distribution of the banks across EU countries. It is seen from the graph that Germany has the highest number of banks amongst EU countries, more than 1,500 banks, followed by Poland in the second place with less than half of the number of German banks. The reason why German has so many banks is that there are hundreds of local, cooperative banks, especially those under Volksbanken-Raiffeisenbanken, operating in cities and villages of Germany, and each of them is registered as an individual bank. This fact makes the banking community in Germany considered as big, in contrary with Netherlands, for instance, which does not follow the same pattern.

As seen from the attack target dataset, the data was sourced from the period between February 2014 and November 2017. It consists of records associated with one of 30 unique threats / malware variants/families in the dataset. The threats available in the dataset were: BokBot, Citadel, CoreBot, Dridex-Loader, Dyre, Gootkit, GootkitLoader, Gozi-EQ, Gozi-ISFB, Ice9, KINS, Kronos, Matrix, NuclearBot, Nymaim, Pkybot, Qadars, Qakbot, Ramnit, Ramnit-BankerModule, ReactorBot, Retefe-v2, TheTrick, Tinba-v1, Tinba-v2, ZeuS, ZeuS-OpenSSL, ZeuS-P2P, Zeus-Action and Zeus-Panda.

By merging the attack target dataset with the list of EU banks, it was found that, out of 5,039 banks, 1,818 banks were targeted, which left 3,221 banks being perceived as not targeted. However there are 1,188 banks on the list whose online banking entities cannot be found. The reason will be provided later in section 5.5. If these 1,188 banks are not considered, it is seen that 1,802 banks were found targeted and 2,049 not targeted. Only 3,851 banks with an online banking service will be taken into account in the further analysis.

5.2. Insights on Distribution of Attack Overtime

One insight that can be extracted from the attack target data is the distribution of attack overtime within the target period (February 2014 - November 2017). Furthermore, the distribution can be created for each country since the country information is clear in each record.

[1] United Kingdom is still considered as part of EU

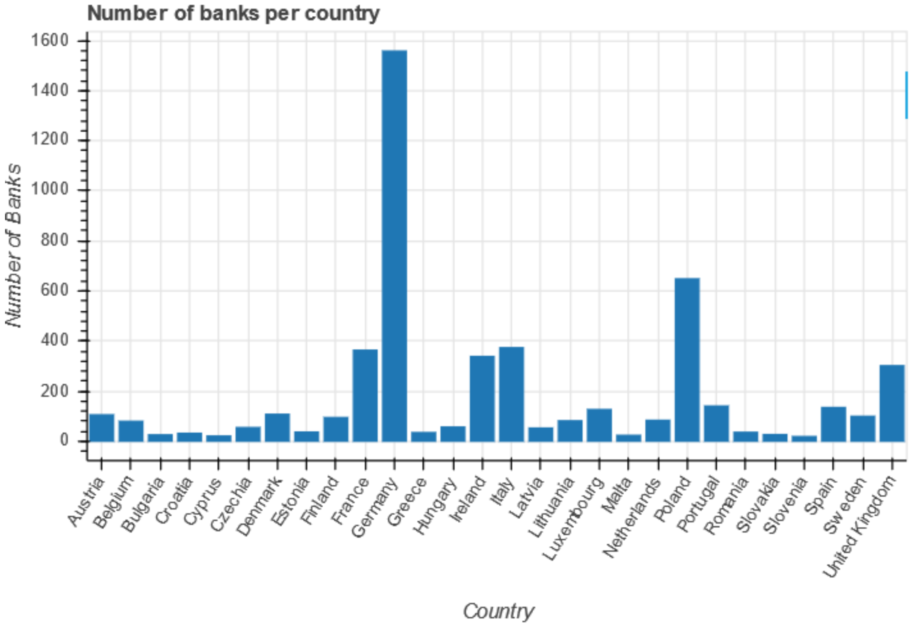


Figure 5.1: Number of banks in every European country

Figure 5.2 plots the attack count during the target period for every country in EU region, based on the attack target data. The attack is counted in raw. There are two interesting findings that can be obtained from the plot. First, it is seen that Germany had, in general, the highest attack count among other countries in EU region. This makes sense since Germany has many more banks than other countries (see Figure 5.1 for number of banks in each country). Second, the notable rise in attack activity can be seen in two periods of time: in April - August 2016 with the peak in May 2016, and September 2016 - January 2017 with the peak in December 2016. Another plot is made to see which type of malware that was responsible for the rise in the activity. Using the same definition of attack count, figure 5.2 also plots the attack count during the same period for every threat / malware variant. It is seen that the rise in a malware activity did not always lead to the rise of attack count received by a country. The characteristic of the attack target URL itself also played a factor. The rise in attack activity in Germany as illustrated in the former graph was caused by the presence of generic URLs inserted in ZeuS-OpenSSL injection code. With such generic URLs, an injection code has a capability to target multiple banks. If the effect of generic URLs are removed, it is seen that three malware variants stood out during the period of analysis: Dyre in 2015, Citadel in 2016 and TheTrick in 2017.

However, as argued in chapter 4, raw attack count might cause an analyst to over-count the actual attack. Therefore, besides the raw attack count, this thesis will also try to make the similar analysis using a proposed attack ID count metric. Figure 5.3 provides the attack ID count during the same period. It is seen that, by using the new metric, Citadel no longer dominated the year 2016. If the metric is proved to be better at counting the actual attack, then it is found that the attack corresponding to this malware variant in that year are overcounted, meaning that there were lots of updates of Citadel configuration files at that period which apparently did not introduce a new attack. Dyre and TheTrick, however, still dominate the year 2015 and 2017, consecutively. This suggests that a lot of new attacks were introduced using such malware variants in those periods. Some external sources are seen to support these findings, like there was an actual surge in Dyre malware infections in 2015 (Carman, 2015) and TrickBot, the other name of TheTrick, was said the busiest financial trojan during the summer of 2017 (Bisson, 2017).

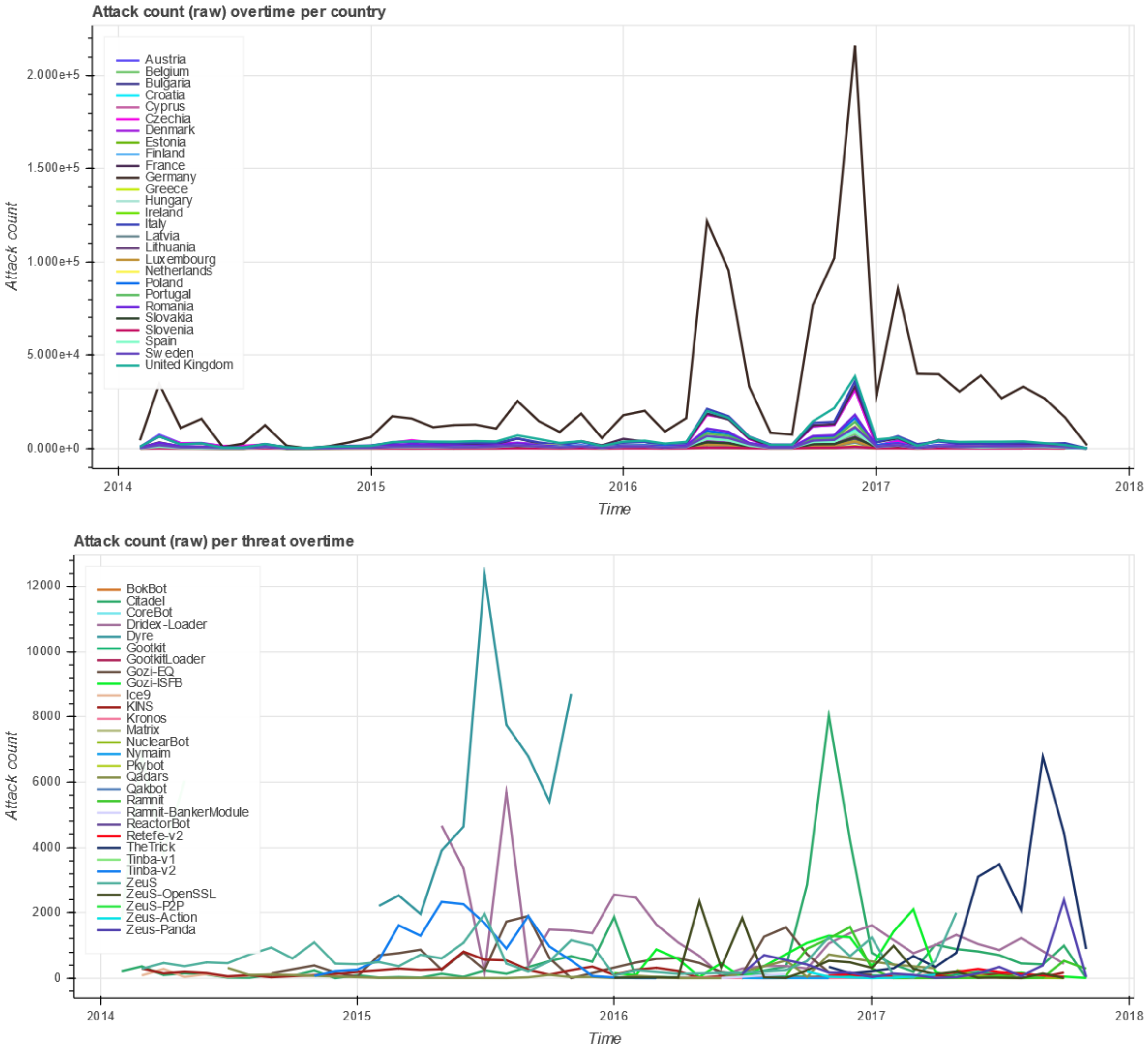


Figure 5.2: Attack Count (raw) overtime per Country (above) and per Threat (below)

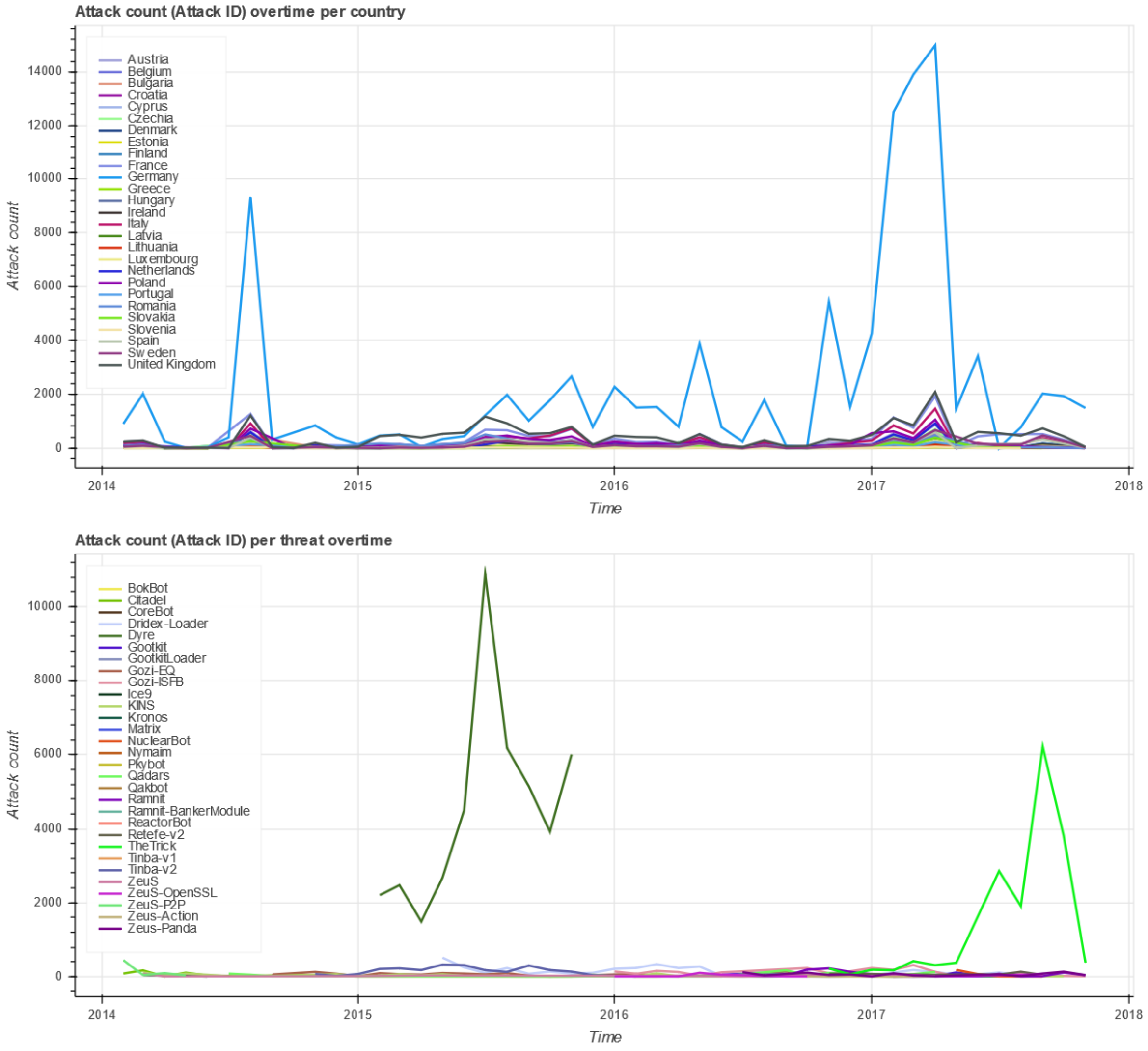


Figure 5.3: Attack Count (attack ID) overtime per Country (above) and per Threat (below)

In this plot, Germany in general maintains its position as the recipient of the most attacks. Different with the former graph, the peak of attack to Germany in 2016 disappeared, indicating that many attacks counted in the previous graph are from the same injection code. A new peak is found in early 2017, caused by Gozi malware. However, a further observation shows that this peak happened due to the presence of generic URLs in the Gozi malware's injection code targeting multiple banks including lots of banks in Germany. Therefore, the high number of German banks could be a reason why Germany got a lot of attacks. The presence of Gozi's new injection code itself is not significant compared to, for instance, TheTrick. This could be an indication that malware variants like Gozi and Zeus-OpenSSL were used to perform more general attack, without really targeting specific entities.

5.3. Insights on Countries Targeted by Different Malware Variants

Another descriptive that can be obtained from the attack target data alone is the relationship between countries and malware variants. The attack target data has already provided the name of malware that corresponds to a certain attack target record. Combined with the country informa-

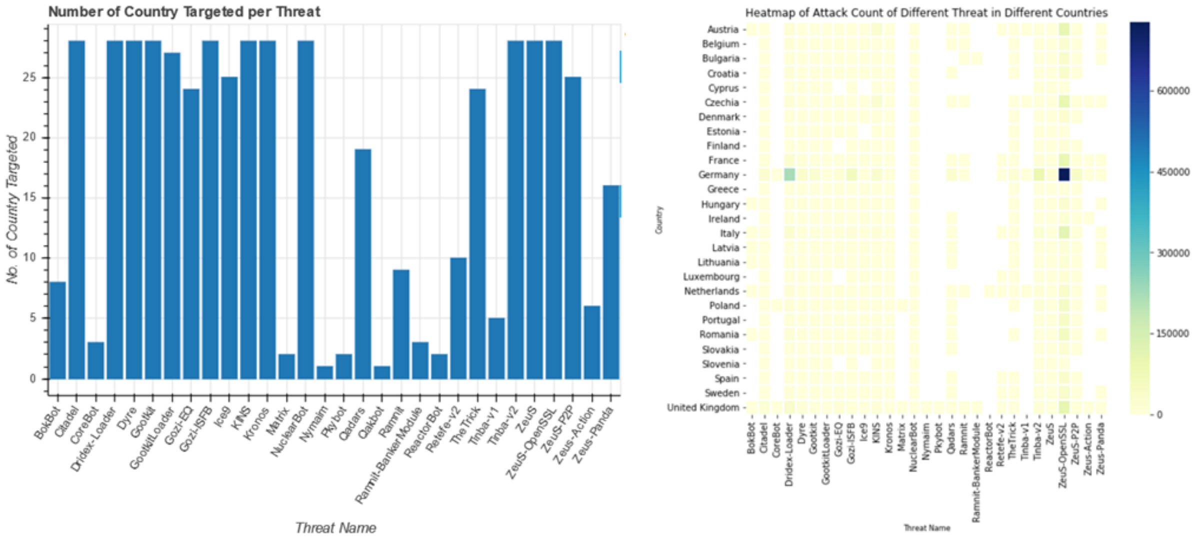


Figure 5.4: A plot on number of country targeted per threat (left) and a heatmap showing the attack count of a threat in a country, based on the attack target data (right)

tion, the insight about which malware targeted which country can be extracted.

Based on the information that some malware were made and used only to attack entities in a particular or several countries, the plot is made to see the number of countries targeted by each malware that is present in the data. The plot can be found in figure 5.4 (left side).

It is seen from the plot that not all malware targeted the same number of country. There are even some malware variants that target only one country according to the data, namely Nymaim, Pkybot and ReactorBot. In order to find out which country they targeted, another plot is made cross-tabulating the threat data and the country data. Figure 5.4 also presents a 2-dimensional heatmap relating the threat to the country. The difference in color in a cell of the heatmap indicates the difference in attack count for a particular threat and country.

The heatmap shows that both Nymaim and PkyBot targeted only the United Kingdom (UK), while ReactorBot only targeted the Netherlands. Further analysis shows that Nymaim only targeted HSBC in the UK, while Pkybot targeted several entities in the UK i.e. The Cooperative Bank, Lyods Bank, Barclays United Kingdom, Halifax and Santander UK. ReactorBot itself only targeted ING Netherlands and Rabobank.

This analysis cannot precisely predict that only those countries and entities are targeted by these threats, as the false negatives from the domain identification process could introduce bias to the analysis result. However, the result still indicates a relatively limited number of countries targeted by these threats and a tendency that these threats are specifically used to perform a targeted attack. This explanation is particularly acceptable for Pkybot and ReactorBot. Pkybot has been identified as a targeted threat focusing on banks in UK, Spain and Greece in 2015 although the web inject for Greece was removed during the financial crisis in that country (Schwarz, 2015). ReactorBot, on the other hand, was first discovered in 2015 and initially targeted banks and financial institutions in Netherlands and Germany, according to Secureworks (2016).

5.4. Insights on the Period of Attack per Threat

Besides counting the attack occurence, the analysis can also be performed for the dimension of time. One of the possible insights is the period of attack per threat within the period of scope. The plot is created by looking at the presence of attack made by a certain threat on a daily basis. Figure 5.5 depicts the period of attack for each threat in the attack target dataset.

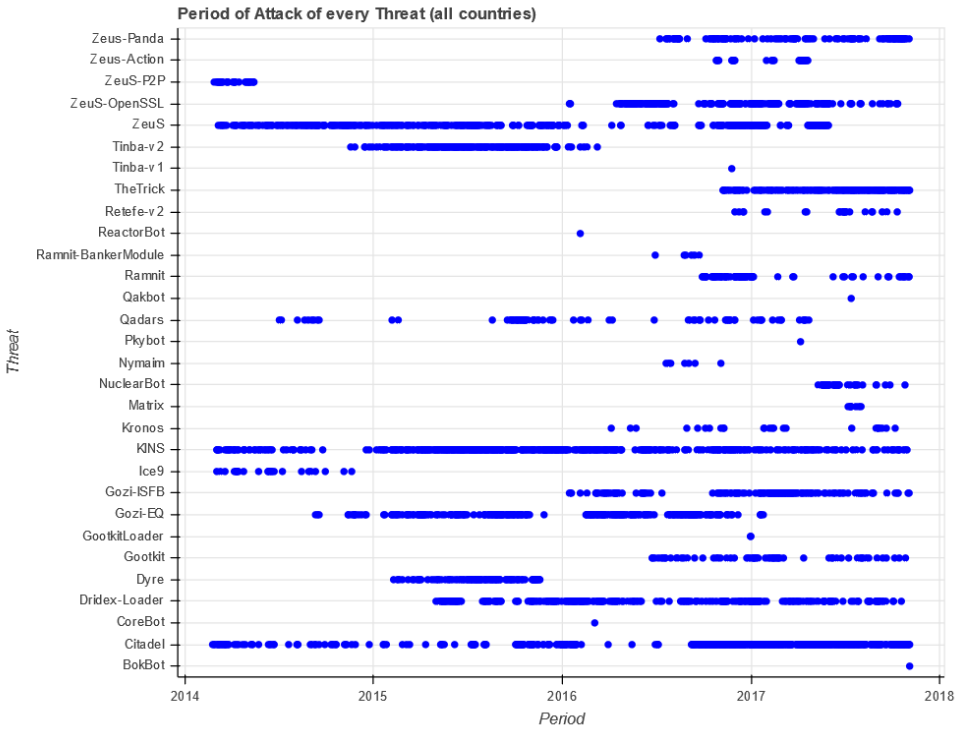


Figure 5.5: Period of Attack per Threat

It is seen from the plot that, while many of the threats are persistent in performing attacks like ZeuS, KINS and Citadel, some of the malware, like ReactorBot and Pkybot, attack at a very specific time. This insight makes these two threats interesting because, in the previous analysis, they were also found to only target a specific country, which suggests a high chance of them being used for specific, targeted attacks.

Also with the dimension of time, this thesis tries to look at the pattern of attack overtime to an online banking entity by a malware variant. There is a hypothesis made by Fox IT that attackers may initially perform small scale attacks to an entity in order to test the effectiveness of the attack, then perform the real attacks after some time of dormancy. This thesis tests this hypothesis by looking at the case of TrickBot attack to ING, a bank in the Netherlands. TrickBot was found to target ING in 2017 (Voolf, Boddy, & Smith, 2017). Attacks to ING Netherlands that correspond to the TrickBot were counted using both raw count and attack ID count approach. They are displayed in figure 5.6. The plot shows an interesting pattern, that is, there are a series of small attacks occuring between March and August 2017 followed by a short dormant period throughout August before a series of relatively intense attacks occuring from September 2017 onwards. Should the hypothesis be true, the small attacks before August 2017 was only trial attacks, while the real attacks were performed from September. Although more analysis must be done in order to get a plausible conclusion about whether this pattern could indicate the actual attack to an online banking entity, this visualisation could become a starting point to analyse whether a bank is really attacked by the malware criminals.

5.5. Insights on Number of Targeted and Non-Targeted Banks per Country

It was possible to gain these insights from merging the attack target data with the banks list. Only countries which are present in both datasets can be visualised, which, in this case, is EU countries.

In this analysis, a bank is considered targeted if there is at least an attack record targeting the entity, which refers to the bank. By merging the list of banks with the attack target data, the attack

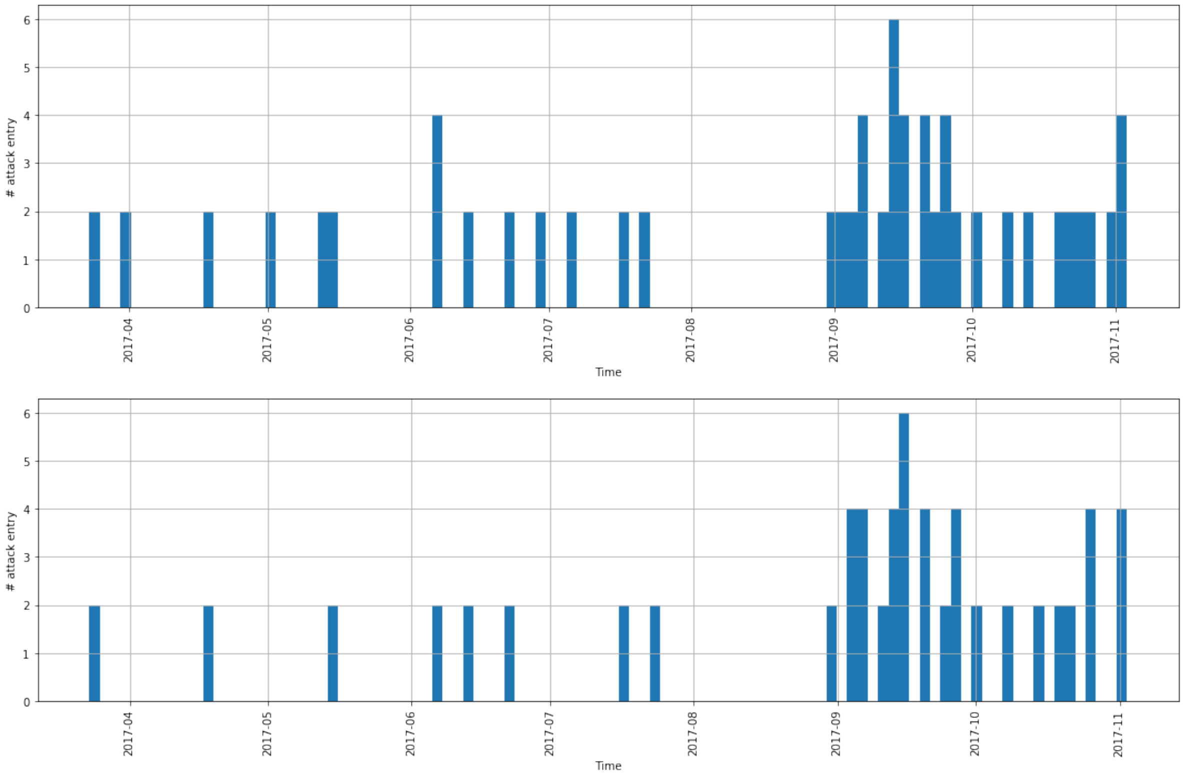


Figure 5.6: Count of TrickBot attack to ING Netherlands, using raw count (above) and attack ID count (below) approach

count for each bank in the list can be obtained, and therefore can be evaluated whether the bank has ever been targeted or not. Figure 5.7 provides the insight on the number of targeted and non-targeted banks per country in EU.

As mentioned in the beginning of this chapter, there are 1,188 banks whose online banking services cannot be found. These banks correspond with either corporate banks, private banks or small, local cooperative banks which seem to only provide traditional channels to their customers like physical address, phone number or email address. Many of them are also not appear in the attack target data. There are, however, 16 banks that are seen in the attack target data although their online banking cannot be found. Looking further at their records in the attack target data, it is seen that their domains is not there any more or redirected to another domain or entity. Attacks to them can be traced back to more than a year. It might be an indication that many things have changed between the period when the attack target data was collected and the period of collecting external parameters such as languages and authentication factor, for example, a bank was acquired by another entity or an online service was probably shut down by a bank. As these banks are outside the scope of attack by banking malware, it is rational to only focus on the banks with the online banking for the further analysis. Therefore, in addition to the previous visualisation, figure 5.7 also provides the insight on the number of targeted and non-targeted online banks per country in EU.

It is seen that Germany had the highest number of banks targeted by the malware. However, this may occur because Germany also has the most banks out of the countries on the list. Yet, still, only less than 50% of the total number of banks in Germany were targeted. This is also true for many other countries. Looking deeper into the merged data, many of the banks not targeted in those countries (especially Germany, Poland, the United Kingdom and Ireland) are cooperative banks. They are banks organized on a cooperative basis, mostly covering only a limited local area in the country. Cooperative banks often have their own online banking services.

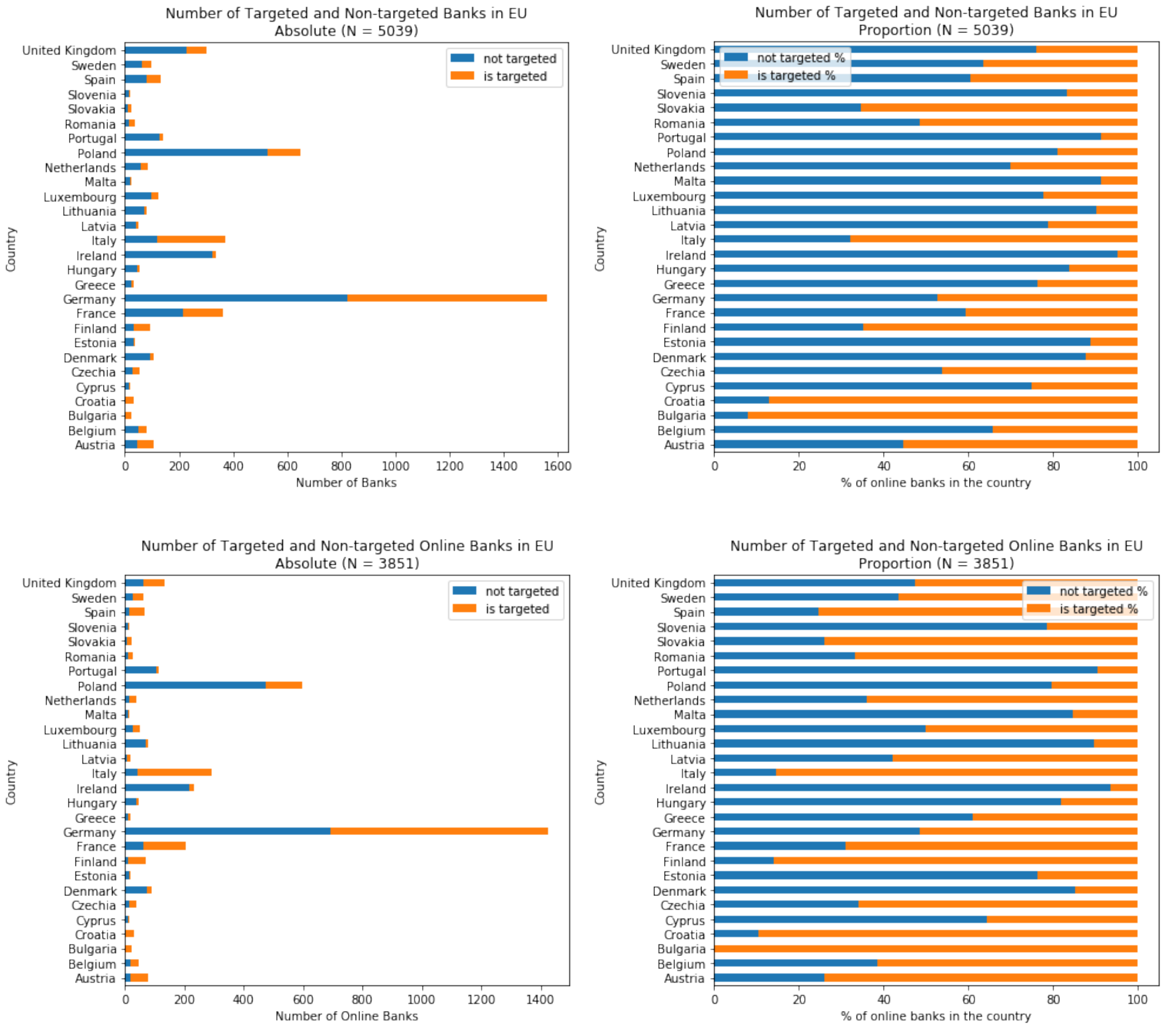


Figure 5.7: Top: plot on number of targeted and non-targeted banks per country in EU (left: in absolute number, right: in proportion). Bottom: plot on number of targeted and non-targeted online banks (banks without online banking are removed) per country in EU (left: in absolute number, right: in proportion)

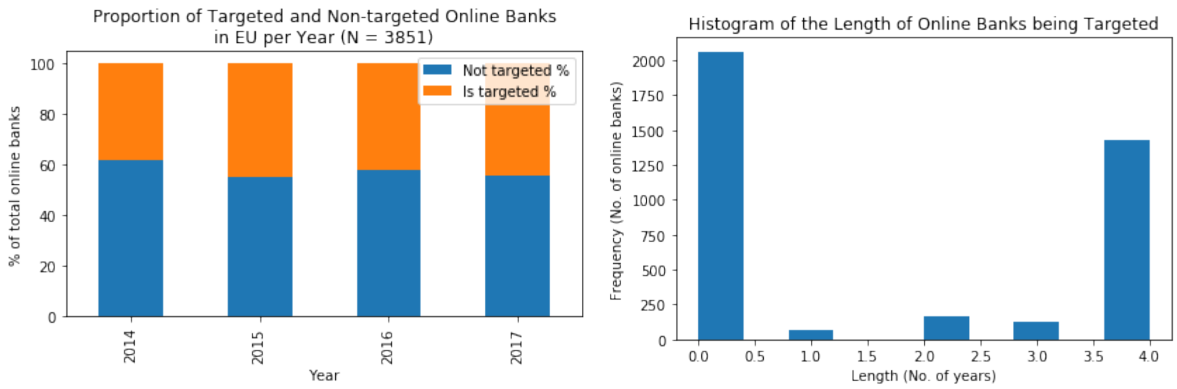


Figure 5.8: Left: Proportion of targeted and non-targeted online banks in EU per year. Right: Histogram of the length of online banks being targeted. It is seen that many of online banking entities were persistently targeted.

Meanwhile, Croatia and Bulgaria have a high ratio of their banks being targeted by the malware, despite only having a few number of banks. It is seen in figure 5.7 that more than 80% of banks in both countries were targeted. Moreover, if the banks without an online banking entity were omitted, all online banks in Bulgaria were seen targeted by banking malware.

If the proportions of targeted and non-targeted online banking entities were evaluated every year, as in figure 5.8, it is seen that the proportions were relatively consistent. The proportion of targeted entities in 2015 was slightly higher than in 2014. Afterwards, the proportion stayed at around 40-45%. This indicates that the criminals had a tendency to target the same victims throughout the year instead of targeting new entities. This indication is also supported by the finding that most of the entities have been targeted for 4 years, as also illustrated in figure 5.8. This means that most of the targeted online banking entities were old victims which were persistently targeted.

The analysis on the targeted and non-targeted online banks continued with looking at how a certain characteristic is distributed across the targeted and non-targeted banks in the EU. In the data collection phase, data about online bank factors were gathered. Through the merging process, these factors can be incorporated to the attack target data in order to get a novel insight.

A characteristic which can be looked into is the popularity of an online banking domain. The top 1 million popular domains were obtained and, using this list, it was possible to determine the proportion of online banks whose domains are on the list versus those not on the list. Figure 5.9 depicts this proportion for the whole EU as well as the breakdown for each country in the area. It is seen that there is a significantly higher proportion of online banks whose domains are on the list for the targeted entities compared to the non-targeted ones. Looking at the whole ratio, targeted online banks have 20% higher proportion of having their domains in the top 1 million list than non-targeted banks. The breakdown per country also shows that, in many countries, the proportion of the presence of online banking domains on the list is higher for the targeted banks than the non-targeted ones. In other words, more popular online banking domains are in the targeted population and it gives a preliminary indication that domain popularity induces a higher chance for an online banking service to be targeted.

The next characteristic to look into is the distribution of the presence of English, that is, whether the online banking service offers English. A plot similar to the one for the previous domain popularity is created and can be seen in figure 5.10. It is seen that the targeted group has a higher proportion of online banking entities offering English. However, the pattern is not as obvious as the popularity of an online banking domain. The trend also varies in every country. In countries like Austria, Croatia, Czechia, Estonia, Greece and Lithuania, a higher number of online banking services offering English is evident for targeted banks. However, other countries do not really show a significant

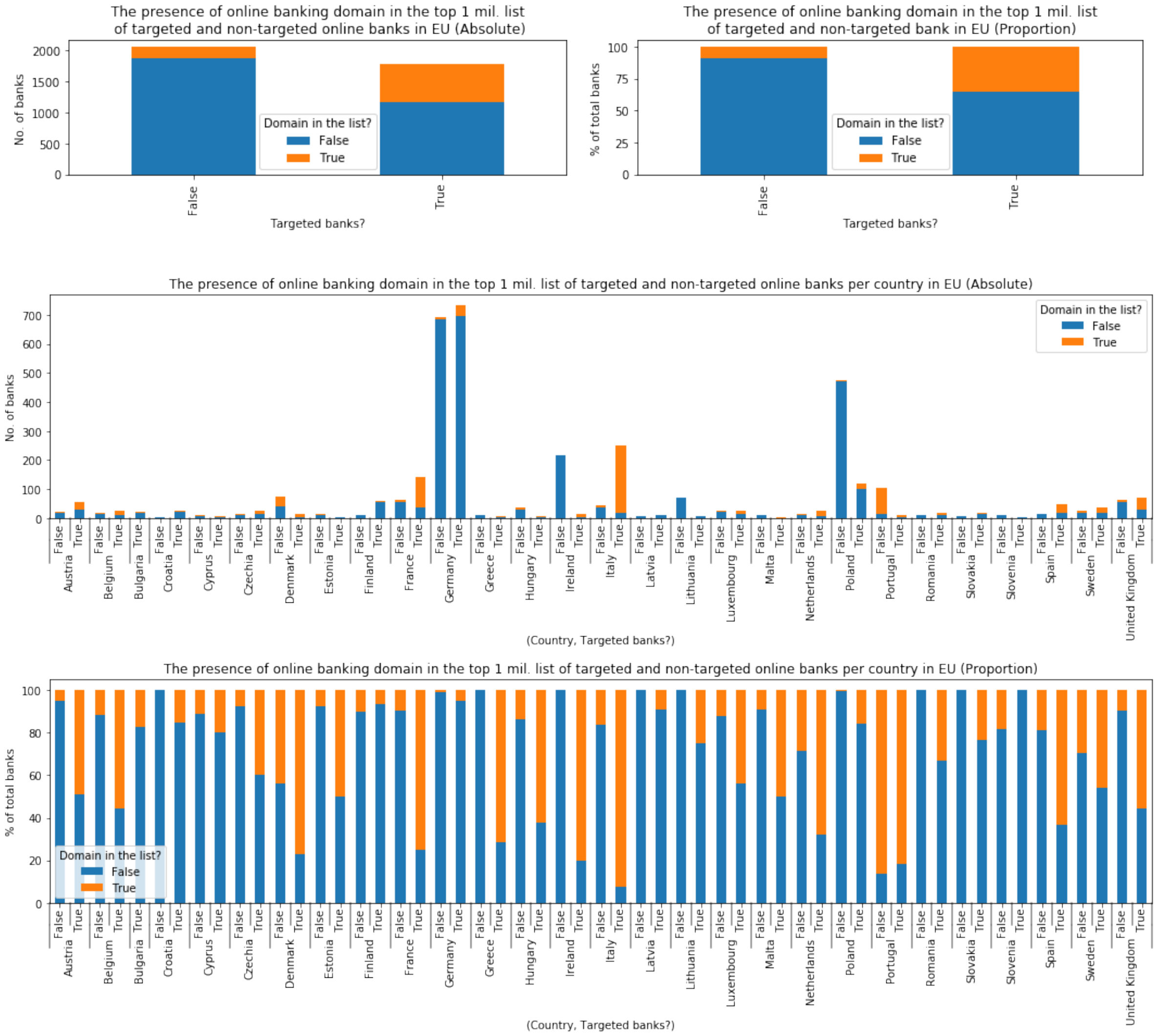


Figure 5.9: Top: proportion of total number of online banking whose domain listed in the top 1 million most popular domain in EU (left: absolute number, right: in proportion. Middle: breakdown of the proportion per country, absolute number. Bottom: breakdown of the proportion per country, in proportion

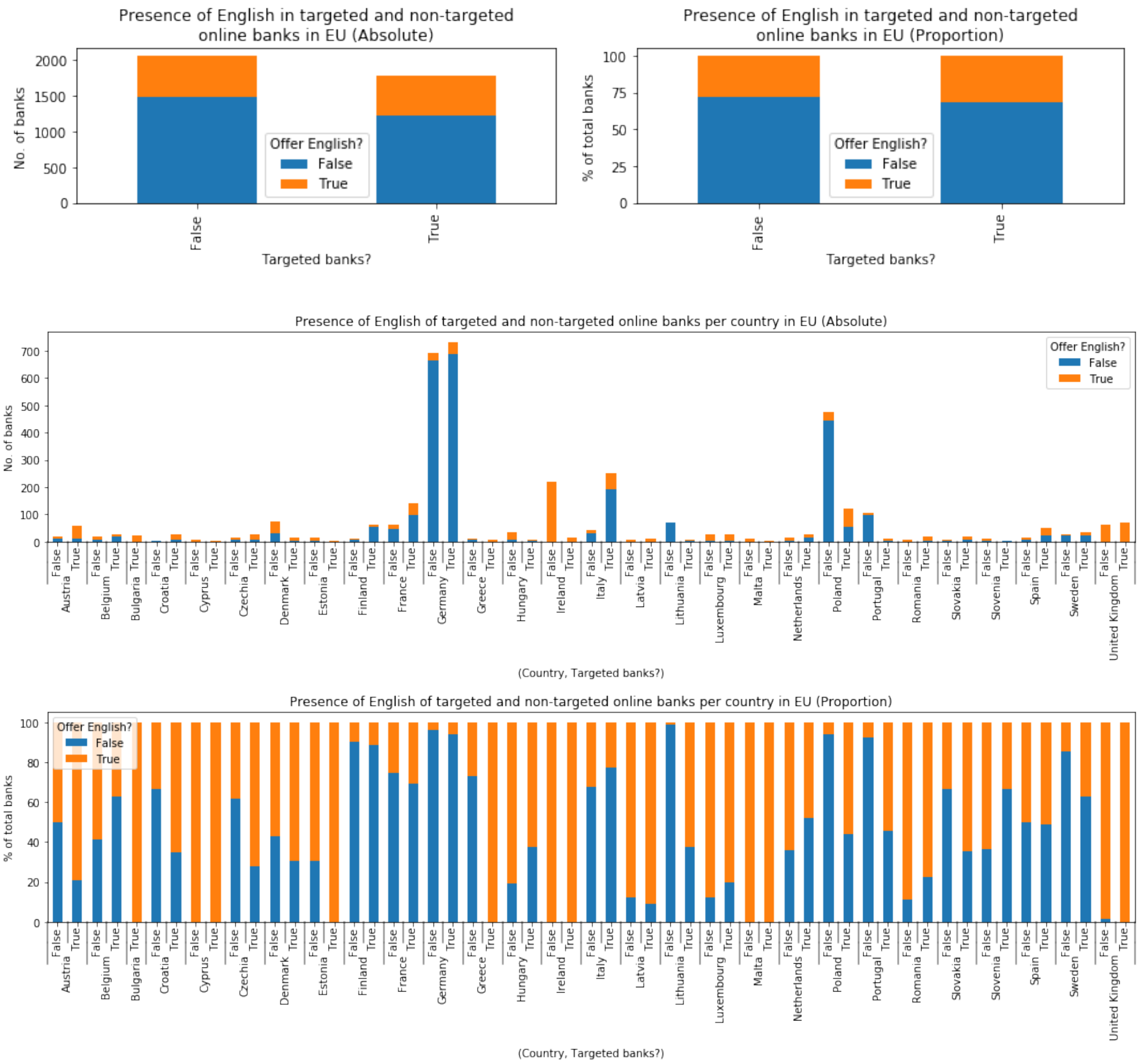


Figure 5.10: Top: proportion of the presence of English of online banks in EU (left: absolute number, right: in proportion. Middle: breakdown of the proportion per country, absolute number. Bottom: breakdown of the proportion per country, in proportion

difference, like the ones found in Finland, France and Germany. Meanwhile, several countries like Belgium and Hungary are seen to be displaying the opposite trend. Therefore, it is difficult to conclude solely from this plot whether English certainly had an influence on the target selection in the whole EU area.

The distribution of the presence of two-factor authentication, as seen in figure 5.11, shows that the proportion of the presence of 2-factor authentication in the targeted groups is higher than the proportion in the non-targeted groups. Similar to English, the difference comes down to the minute. Furthermore, the trend of this proportion is also different for every country.

5.6. Concluding Remarks

This chapter shows that there are a lot of insights that can be extracted from the datasets which were collected. These insights are useful to gain further understanding about the attack landscape as well as to get a preliminary sense of the target selection of the banking malware.

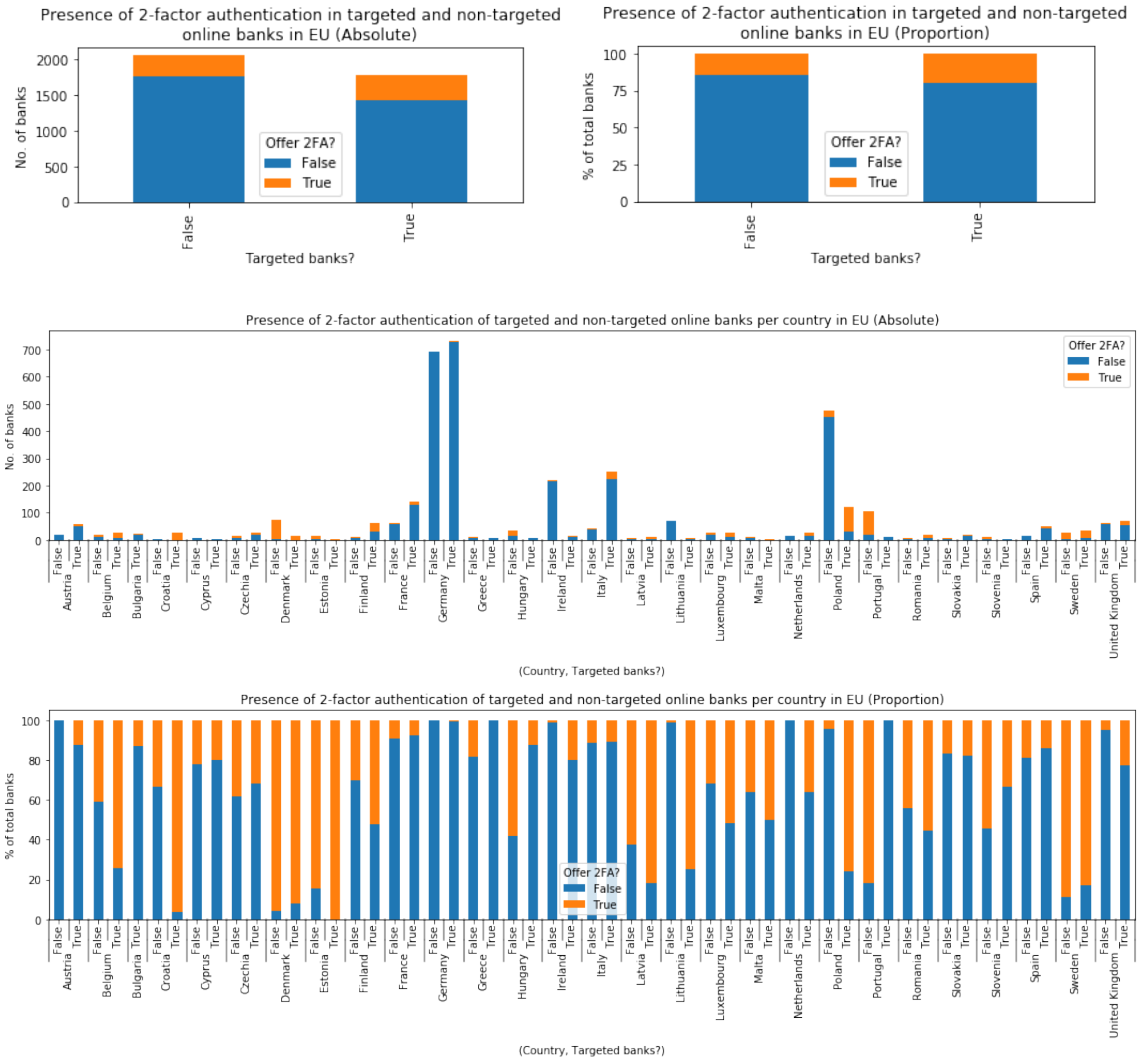


Figure 5.11: Top: proportion of the presence of 2-factor authentication of online banks in EU (left: absolute number, right: in proportion. Middle: breakdown of the proportion per country, absolute number. Bottom: breakdown of the proportion per country, in proportion

Recalling the findings from this chapter, it is found that different types of malware targeted different entities in different countries and had different periods of attack. Some malware variants may be utilised for performing persistent attacks and targeting wider areas, like ZeuS, KINS and Citadel, while other variants may only be used to attack a specific target at a specific time, like ReactorBot and Pkybot. This information is useful for categorizing the malware based on whether the malware is designed to perform a generic attack or a specialized attack.

The analysis on the targeted versus the non-targeted online banking entities showed that the proportion of targeted and non-targeted entities is different per country. For example, Bulgaria becomes a country in which nearly all of its banks were targeted by the malware. On the other hand, Germany, while having the highest number of banks targeted by the malware, is not the country with the highest proportion because Germany is home to a large number of banks where many of them were not targeted.

It is also shown in the further analysis that the targeted group has a higher proportion of online banking entities being in the top 1 million of popular domains as opposed to the non-targeted group. This validates the finding that, in general, criminals tend to target popular online banking entities. On the other hand, even though the presence of English and 2-factor authentication is higher in the targeted group than the non-targeted group, the difference is only little. Therefore, it is difficult to conclude anything from these two characteristics for now. The explanatory analysis will be performed to see further how these characteristics relate to the target selection.

6
Explanatory Analysis

Explanatory analysis is conducted in this thesis in order to find the relationship between characteristics that have been identified and selected before and the target selection. The analysis will be done by making a regression model. Before the model can be created, it is important to identify variables that will be considered in the model.

6.1. Variables for Explanatory Analysis

This section explicates both dependent/response and independent/predictor variables that will be included in the explanatory analysis. The variables are based on metrics that have been extracted earlier in this thesis.

6.1.1. Dependent Variables

This thesis will try to find out whether the selected characteristics, i.e. language and authentication factor, of online banking entities could explain the target selection. Therefore, metrics that correspond to the attack count will become the dependent variable for the model.

Earlier in chapter 4, three metrics of counting the attacks were chosen: the raw attack count, the week-interval attack count and the unique attack ID count. This thesis will try to generate a model for each approach. As a consequence, three different models will be generated and presented.

6.1.2. Independent Variables

Independent variables used for the model are divided into variables to be assessed and variables that serve as the control variables, so that their effects can be separated. The variables to be assessed are listed below:

The presence of language

Multiple dummy variables are created from the language data of the banks on the list to make clear whether a certain language is offered by the online banking service. A dummy variable of a language will contain a true value if the language is offered and a false value if the language is not offered.

The created variables cover languages that are present in the EU. They are: English, German, French, Dutch, Italian, Spanish, Portugese, Greek, Czech, Slovak, Slovenian, Polish, Hungarian, Romanian, Bulgarian, Danish, Swedish, Finnish, Latvian, Estonian and Lithuanian.

Number of languages offered (language count)

A count for the number of languages offered by an online banking.

The presence of authentication factor

Similar to the presence of language variables, dummy variables for authentication factor are created from the authentication factor data of the banks on the list. These variables show whether a 1-factor authentication is offered and/or 2-factor authentication is offered.

Meanwhile, the following variables will be the control variables:

Country

The country where the bank is located, according to the list of banks from European Central Bank (ECB).

Threat

The malware variant that attacks the target as indicated in the attack count.

Year

The year when the target was attacked as indicated in the attack count.

Unique URL count

The number of unique URLs corresponding to the online banking entity that were present in the attack target data.

Popularity score of the bank's domain

This variable relates to the domain popularity of the online banking site. However, since many banks' online banking domains are included in the top 1 million list, there are many missing values in this variable. These missing values may reduce the quality of the model because they make many data points unusable, causing the generated model to fit with fewer, more unrepresentable data points.

In order to address this problem, a "zero-coding" data imputation technique, which assigns zero values to the missing values, is utilised (Gelman & Hill, 2006). This imputation works for an inverse ranking, that is, the first rank on the list is given the highest score proportional to the total number of items on the list; in this case a million, and then the scores decrease for the next ranks in the same interval as the original rank. Those who are not part of the rank, meaning they are not in the top 1 million, gets the score of 0. Although this imputation reduces the sensitivity of the results, it provides a simple transformation to improve the predictive power of the regression model (Gelman & Hill, 2006).

6.2. Data Preparation

The data must be made ready to be used in the building the model. This is done by locating and removing data points with not available (NA) values.

Although many NA values have been imputed, as in the case of domain popularity score variable, NA values are still present not at random. This corresponds to 1,188 banks whose online banking cannot be found, as explained in section 5.5. Therefore, they are not included in the data for regression, leaving 3,851 entities used for model generation. Moreover, since the threat/malware variant and year factor are included for the model, the dataset was adjusted so that each datapoint refer to the attack count to an entity by a malware variant in a year. As a result, a dataset of 35,471 datapoints is ready for the analysis, with 33,384 datapoints of targeted entities.

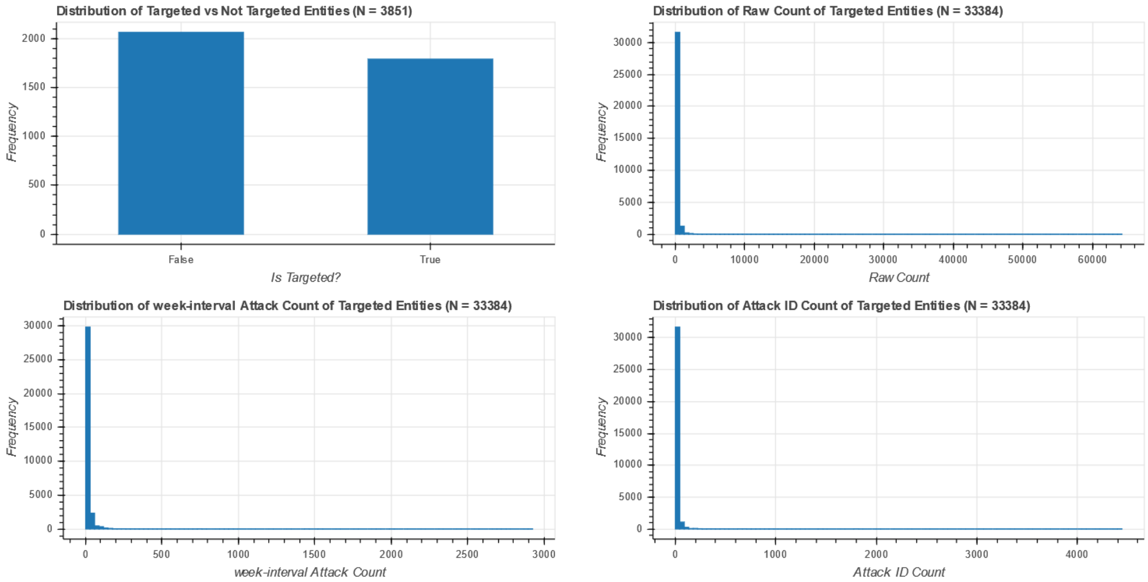


Figure 6.1: Distribution of response variables. Top left: frequency of targeted vs non-targeted entities. Top right: distribution of raw attack count of targeted entities. Bottom left: distribution of week-interval attack count of targeted entities. Bottom right: distribution of attack ID count of targeted entities.

6.3. Data Distribution

The distribution of dependent variables were analysed to find out the type of their data distribution and the suitable regression model for the analysis. Figure 6.1 displays the distribution of the dependent variables. It is seen that the distribution for every attack count metric follows a negative binomial distribution where the standard deviation is higher than the mean.

This thesis will approach the regression analysis with two types of regression models. First, logistic regression is used to model the probability of being targeted (having an attack count more than zero) or not targeted (having an attack count of zero). The logistic regression is useful to address potential issues caused by a lot of zero values in the data; singularity problem for instance. Second, for the targeted entities, the data will be fitted by using a Poisson-family model, which is suitable for a dispersed count. The first attempt is to get the dispersion parameter of the data in order to justify which model is more suitable. Poisson GLM was run to check the dispersion parameter of raw attack count, 7-day interval attack count and the unique attack ID count. It is found that the dispersion parameters for all of them are 323.98, 36.80 and 20.26, respectively. The dispersion parameters are significantly higher than 1, indicating that the data is over-dispersed, meaning that there is "unobserved heterogenity in terms of a missing structural factor that leads to concentrations of observable events" (Tajalizadehkhoob, Böhme, Gañán, Korczynski, & van Eeten, 2017, p. 11). This leads to the use of negative binomial regression, which is more suitable for an over-dispersed count data distribution. The use of any zero-inflated regression model is discouraged in this case since the number of zero values is still not excessive enough.

6.4. Regression Model and Results

As explained previously, two types of regression model will be generated: logistic regression which will fit the data to explain the probability of an online banking to be targeted (attack count > 0) or not be targeted (attack count = 0) and negative binomial regression which will fit the data of the targeted entities to explain the tendency of an online banking entity being more or less targeted. The negative binomial regression will be fitted to all attack count metrics.

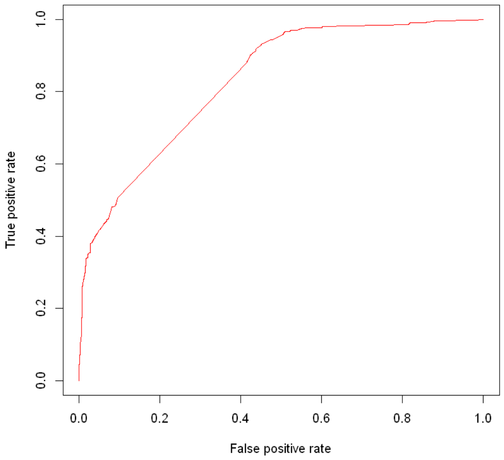


Figure 6.2: ROC curve for logistic regression

The first model is the logistic regression. It is made to explain the probability of an online banking entity being targeted or not targeted. A new variable is created whether an entity is targeted or not; true (1) if the attack count is more than zero for the entity and false (0) if the attack count is zero. This new variable becomes the dependent variable for the model. For the logistic regression, the presence of a particular language, the language count, the presence of an authentication factor, the domain popularity score and country are included as predictors. The other variables extracted from the attack target data are not included because they do not provide the variability required for performing the regression, for example, the number of unique threats attacking an online banking entity is also zero if the raw attack count is zero.

The model has a difference of 1,627.504 between null and residual deviance, with the associated p-value of 2.65e-307 (below 0.001), which indicates that the model with predictors is significantly better than the similar model with only intercept (null model). The model also has McFadden R-square value of 0.306. Further analysis using this model showed that around 70% of data points were correctly categorized, while around 25% were identified as false negative (actually targeted but not identified as targeted) and 5% as false positive (identified as targeted while actually not targeted). From the ROC curve, shown in Figure 6.2, it is seen that the area under curve (AUC) is around 0.831.

The second model is negative binomial regression, which is performed for data points which have the raw attack count of more than 0. Besides the predictors used in the logistic regression, other predictors that are extracted from attack data, such as threat, threat count and number of unique URLs, are included. The model has a difference of 82,806.56 between null and residual deviance, with the associated p-value of 0 (below 0.001), which indicates that the model with predictors is significantly better than the similar model with only intercept (null model). The model also has McFadden R-square value of 0.131.

The third model is the negative binomial regression which is performed towards week-interval attack count. The predictors used are the same as those for the regression towards raw attack count. The model has a difference of 52,671.1 between null and residual deviance, with the associated p-value of 0 (below 0.001), which indicates that the model with predictors is significantly better than the similar model with only intercept (null model). The model also has McFadden R-square value of 0.135.

The fourth model is the negative binomial regression which is performed for the attack ID count. The predictors used are the same as those for the regression towards raw attack count and 7-day interval attack count. The model has a difference of 87,265.09 between null and residual deviance, with the associated p-value of 0 (below 0.001), which indicates that the model with predictors is significantly better than the similar model with only intercept (null model). The model also has McFadden R-square value of 0.199.

As an illustration, the regression result for several language and authentication factor variables are presented in table 6.1. More detailed outcome and summary of the above regression models can be seen in appendix B.

Table 6.1: Short summary of regression models, for several variables

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
Lang: English	0.629* (0.314)	0.379*** (0.029)	0.124*** (0.025)	0.193*** (0.025)
Lang: German	0.515 (0.385)	0.265*** (0.038)	−0.150*** (0.032)	0.005 (0.033)
Lang: Dutch	0.604 (0.561)	0.277*** (0.060)	0.364*** (0.051)	0.396*** (0.052)
Lang: Italian	0.847 (0.639)	−0.046 (0.058)	−0.279*** (0.049)	−0.256*** (0.050)
Lang: Spanish	−0.061 (0.625)	0.947*** (0.072)	0.537*** (0.061)	0.794*** (0.061)
Lang: Portugese	0.885 (1.022)	−0.228* (0.089)	−0.677*** (0.077)	−0.868*** (0.079)
Lang: Swedish	2.146** (0.793)	0.490*** (0.138)	0.318** (0.118)	0.379** (0.118)
Lang: Estonian	−3.439*** (0.916)	0.289* (0.135)	−0.557*** (0.119)	−0.249* (0.123)
2-factor Auth.	2.148*** (0.266)	−0.207*** (0.033)	−0.173*** (0.028)	−0.106*** (0.028)
Note:				*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets

6.5. Reflection on the Models and Results

Based on the model result, it is seen that the language variables, in general, do not show their significance in the logistic regression model. Only English, Swedish and Estonian are seen significant enough. However, the significance of Estonian should be treated with caution as this variable has a strong correlation with the 'Estonia' country variable and seen to be singular in the model. As multicollinearity exists between these variables, it is hard to determine whether the significance happened because of the Estonian language or because the banks are located in Estonia. Meanwhile, it is seen in English and Swedish that the coefficient is positive, which indicates that online banking entities which offer these languages have higher chance to be targeted.

More language variables are seen to be significant in the negative binomial models, indicating that the language may be able to explain why some banks are more or less targeted than others. English is again significant and has a positive coefficient, which indicates that its presence may cause the tendency of the online banking entity to be more targeted. This effect is also relevant for several other languages like Dutch and Spanish. The opposite effect seems to be present for Italian and Portugese. Based on the regression models above, despite the presence of control variables like the threat/malware variants, country, year and domain popularity, some languages are seen to still maintain their significance in explaining whether an online banking entity is more or less targeted.

It is important to note early that the result between models can be different, depending on the attack count metric used. This signals that the way the attack is counted affects the result and hence the quality of the analysis. The closer the metric to counting the actual attack is, the more acceptable the analysis. Although every metric approaching the actual attack count has its own advantages and limitations, this thesis evaluates that the metric that uses unique attack ID to count the attack seems to approach the actual attack count the closest among other metrics. This is due to the metric being based on the legitimate explanation of how the malware data is stored instead of arguable assumptions that other metrics rely on.

It is also seen in the models that the presence of two-factor authentication (2FA) can explain the target selection of banking malware. This variable is seen to be significant in both logistic regression and negative binomial models. This finding seems to contradict Van Moorsel's (2016) argument that financial institutions with two-factor authentication are selected as much over the years as ones with one-factor authentication. However, the interpretation of this variable is quite complicated. The variable has a positive coefficient in the logistic model, which means there are more banks implementing 2FA that are targeted (than banks which do not implement 2FA). On the contrary, it has a negative coefficient in all negative binomial models, which indicates that the presence of 2FA reduces the tendency of online banking to be more targeted. A possible interpretation to these findings is, since 2FA is becoming common nowadays, many criminals are still trying to attack the online banking. However, they may not or cannot perform a lot of attack, or maybe they tried to attack it in the first time and realized that the benefit was unsatisfactory. As the result, it discourages them to perform more attacks, causing a tendency the targeted bank to be less targeted.

In order to see how the significance level of variables changes as more factors were included in the model, several models were created out of the negative binomial regression model towards unique attack ID count. These models took into account different number of predictors: (1) the first model only considers language and authentication factor, (2) the second model adds unique attack URL count to the first model, (3) the third model add domain popularity to the second model, (4) the fourth model adds country variable to the third model, (5) the fifth model includes threat variable to the fourth model, and (6) the sixth model includes year variables, makes it similar to the model initially generated. Some independent variables are selected in order to show how their coefficient and standard deviation change overtime due to the addition of control variables. The summary is presented in table 6.2.

It is seen that the coefficient of the variables changed while more factors are taken into account. However, some variables were able to maintain their significance despite the presence of more factors, like the presence of English and 2-factor authentication. There are also some language variables which lose their significance after the addition of control variables, like Danish. This indicates that some factors might be more important than others. It is also seen that the model gets better in explaining the variance of the data as more factors are included, indicated by lower Akaike Information Criterion (AIC). However, the AIC is still high. Referring back to the initial logistic and negative

Table 6.2: Negative binomial regression towards unique attack ID count

	Response Variable: Attack ID Count					
	Negative Binomial					
	(1)	(2)	(3)	(4)	(5)	(6)
Lang: English	1.108^{***}	0.580^{***}	0.522^{***}	0.407^{***}	0.188^{***}	0.193^{***}
	(0.032)	(0.031)	(0.031)	(0.035)	(0.026)	(0.025)
Lang: Dutch	1.031^{***}	0.496^{***}	0.460^{***}	0.586^{***}	0.392^{***}	0.396^{***}
	(0.044)	(0.043)	(0.043)	(0.072)	(0.052)	(0.052)
Lang: Spanish	1.735^{***}	1.133^{***}	1.119^{***}	1.222^{***}	0.776^{***}	0.794^{***}
	(0.065)	(0.062)	(0.062)	(0.087)	(0.061)	(0.061)
Lang: Danish	0.614^{***}	0.618^{***}	0.542^{***}	0.410^{*}	0.039	0.045
	(0.091)	(0.087)	(0.087)	(0.182)	(0.130)	(0.128)
2-factor Auth.	−0.043	-0.157^{***}	-0.134^{***}	-0.091^{*}	-0.108^{***}	-0.106^{***}
	(0.041)	(0.039)	(0.039)	(0.040)	(0.028)	(0.028)
Unique Attack URL count		0.019^{***}	0.016^{***}	0.016^{***}	0.012^{***}	0.012^{***}
		(0.0003)	(0.0003)	(0.0004)	(0.0002)	(0.0002)
Domain popularity			0.00000^{***}	0.00000^{***}	0.00000^{***}	0.00000^{***}
			(0.00000)	(0.00000)	(0.00000)	(0.00000)
Country				True	True	True
Threat					True	True
Year						True
Log Likelihood	−111,203.800	−109,154.600	−109,084.700	−108,967.800	−93,033.010	−92,708.800
θ	0.634^{***}	0.703^{***}	0.706^{***}	0.710^{***}	1.729^{***}	1.767^{***}
	(0.005)	(0.005)	(0.005)	(0.005)	(0.015)	(0.016)
Akaike Inf. Crit.	222,457.500	218,363.200	218,225.300	218,043.600	186,232.000	185,589.600

Note: *p<0.05; **p<0.01; ***p<0.001

Standard errors in brackets

binomial models, the pseudo McFadden-R^2 of the models is considered low, around 0.3 for the logistic regression and 0.1 for the negative binomial regression. Although pseudo R^2 does not indicate the ability of the model to explain the data variance, like the real R^2, having a low pseudo R^2 suggests that the model is still far from the perfect fit and there might be many other factors out there which could improve the model in explaining the target selection.

The above inference is made based on the generated model and result, and it should be taken into account that the model itself has some limitations that may prevent plausible conclusions to be drawn. The models are still not able to include many more factors due to limited data and time. As an implication, the model cannot explain the influence of other factors that are not considered in the model on the target selection. Some non-considered factors may be more important than the factors considered in this model. Moreover, data quality could also be an issue. As explained previously, the model relies on attack target data from Fox IT and also Fox IT's mechanism in identifying attack URLs which are associated with EU banks. Although the improvement was applied, the unidentified bank-related attack URLs, or the false negatives, may still be present and they may cause the model to not be perfectly accurate. If the identification quality were better, the data would have been more robust and hence the model and the result would have been different to the current ones. Human error may have also occurred during manual observation of online banking entities, which would have affected the data quality.

7 Expert Interviews

This chapter will describe about the final phase of this thesis, the expert interview.

7.1. Objective of the Interviews

The main objective of the expert interviews was to obtain experts' general perception of the target selection and of the factors that may influence it. Moreover, this interviews were also intended to obtain experts' perception and interpretation of the model and of the results which have been generated in the previous phases of this study.

7.2. Overview of the Interviews

The interviews were conducted in a semi-structured manner, so that the discussions could flow naturally and more information and opinion from the experts can be gathered. In order to obtain richer and more objective insights, the interviews were conducted with three experts. The experts are security experts working in the leading banks in the Netherlands, so it is assured that they are very knowledgeable about the target selection topic. The interview protocol and the profiles of the experts are elaborated in the following subsections.

7.2.1. Interview Protocol

The interview consisted of two phases. In the first phase, the questions about their opinion on the target selection were asked. In the discussion, experts explained both financial and non-financial potential factors which they presume could affect the target selection. A slot was also given for them to argue about how the factors that become the highlights in this thesis, languages, authentication factors and domain popularity, could affect the target selection. The first phase was conducted before the model and the results were presented. This was done intentionally so that the experts' perspective was not affected by the model and the result.

In the second phase, the model and result were presented. Afterwards, questions were given to the experts in order to get their opinion about the model and the result. Their interpretation of the outcome of the model were also requested.

The interview protocol document which contains the questions given to the experts can be seen in Appendix C. Since it is a semi-structured interview, the document only serves as a guide and thus the actual questions may deviate from it, but will still be within the context of the established interview protocol.

7.2.2. Profile of the Experts

The interviewees are security experts working in leading banks in the Netherlands. Therefore, they possess good knowledge and experience about the landscape of cyber security in the banking sector. The profile of each expert is as follows.

Paul Samwel

Paul Samwel is the head of security architecture, innovation and cyber threat intelligence division of Rabobank Netherlands. He is a member of MALPAY team, a collaboration team consisting of Fox IT, TU Delft, other banks' representatives and the Dutch police that discusses cybersecurity updates and issues in the banking sector in the Netherlands.

Maarten Jak

Maarten Jak is a member of security and intelligence management team in ABN Amro Netherlands. He obtained a master's degree in System Engineering, Policy Analysis and Management from TU Delft and a minor in Safety, Security and Justice. He has been spending his 6-year career in ABN Amro working on security and integrity management on cybercrime and threat intelligence. He is also a member of the MALPAY team.

Huub Roem

Huub Roem is a forensic IT expert at the cyber defense center within ING Bank. His team serves domestic banks in the Netherlands investigating e-frauds and doing incident response in case of internal threats.

7.3. Output of the Interviews

This section summarizes relevant arguments that experts expressed in the interview sessions. A more detailed outcome of the interview can be found in the transcriptions in Appendix D.

7.3.1. Experts' Perspective on the Target Selection of Banking Malware

An expert expressed that the most important factor to select a target for criminals is something that the criminals can gain. According to his perspective, the first thought criminals will have is whether there is a positive business case for them to target any entity. Financial and market profile of the banks may contribute to the calculation of the business case. He illustrated how the market could make a pattern of the target selection using the banking sector in the Netherlands. The Netherlands has a small banking community, where the top 3 banks account for 90% of the banking market share of the country. This increases the tendency for criminals to focus their attacks in the Netherlands on these banks and not other Dutch banks. Assume that a criminal owns a botnet and a malware that can target one of these top banks, then the criminal would be able to utilise the botnet much better than if the malware target other banks. On the other hand, in Germany, it is harder for a criminal to pick their targets due to the existence of many smaller banks.

After the business case, the criminals will inspect the easier target. This is where factors like the language and authentication factor come into play. An expert expressed that, due to the presence of detection system in banks, the malware is shifted from the automated malware, which blindly scans many online banks and their features for attacks, to more personalised malware involving lots of manpower. It means criminals now tend to make specific malware for specific targeted banks. That is why, according to him, factors like language and authentication factor will come into play. Another expert argued that criminals are acting more with social engineering using social media and other kinds of channels outside the banking channel.

Experts also see security control as an important aspect. An example of this is the quality of detection systems implemented by a bank. An expert saw that the criminals shifted their focus to another country by the time the expert's institution started running security monitoring and became stricter with the money mules by exchanging account information between other Dutch banks and initiating awareness campaigns.

Other factors may also affect the target selection. For example, an expert mentioned the record of successful attack attempts, that is, a successful attack on a bank may inspire the criminals to attack the same bank again with the hope of the same success. The expert also highlighted that the number of banks in the specific area/country would affect the target selection. The argument is that if there is a large number of banks in a country, for example Germany with hundreds of small banks in the country, it will be harder for criminals to attack many victims in one attempt as there are multiple banks a customer may have an account with. Therefore, criminals become unsure whether their efforts in attacking lots of banks there will pay off or not.

The maturity of online banking system of the country and the adoption rate of online banking by the society may also influence the target selection. The expert mentioned the Netherlands as a country in the EU which appears to be more mature compared to others in terms of implementing online banking system. Almost all Dutch banks are now offering online banking services. Dutch society is already accustomed to online transactions and the usage rate of online banking services among Dutch people is considered high. Another potential factor mentioned by an expert is the rate of system change. His argument is, if an online banking service performs relevant changes or updates for avoiding the malware in a quick pace, the malware which is designed to attack it must also be adjusted frequently, which may discourage the criminals.

However, an expert insisted that financial factors still play more important role than the other factors. He argued that, in the end, the criminals will only calculate the financial aspects of the non-financial factors. The expert gave an example about the Dutch tax institute. It had certain authentication factor and security control that were fine for years and saw no fraud in its online service. When it started to offer an online service for claiming subsidy, however, that was when it started to see frauds, despite having no changes in its authentication factor and security control. This example shows that financial traits, which flourished due to the update in the target's service (from only collecting tax report to also providing subsidy), outweigh the non-financial factors like the authentication and security control of the service.

Regarding the language, experts argued that it has a role in the target selection. A foreign language could make the creation of a web inject more complicated because the inject code needs to possess a good command of that language without any typo. English-speaking countries are attacked more, in experts' opinion, as English is more widely used and many people are able to speak the language. An expert believes that due to this reason, criminals look at the English page. However, he is also aware that malware becomes more local nowadays, as criminals start to hire people to translate the phising page into local languages. Experts think that a google-translated phising page is no longer adequate to convince people since customers are getting smarter in assessing a phising page by its language. Therefore, they argued that criminals nowadays hire local people who are able to speak the language of the banks they want to target.

In terms of the German language, an expert argued that since many malware originate from East Europe, it has the same position as a foreign language as criminals are not really familiar with it. However, since there are lots of German-speaking people in Europe Germany possesses a lot of money, it would be easier to recruit a criminal who speaks German. This could explain why German banks are also attractive to criminals. Another expert linked this issue with the fact that there are many small banks registered in german-speaking countries. He argued that it is difficult to write many malware codes to attack many small banks. However, he also highlighted that the malware attack is now getting less automated and more people intensive. Therefore, targeting a big nation with a huge number of banks is not a problem anymore for criminals since they do not have to write an automated transaction code like what they did to the automated malware. He also mentioned that Germany is also not as developed in terms of its online banking as the Netherlands. He emphasized that criminals tend to look at the size of banks and quantity of banks in the region rather than the language they use. Therefore, the language may not be a sole factor which explains why a certain phenomenon regarding the target selection may occur.

Regarding the authentication factor, experts believe that a good form of authentication is a good way to make an attack less successful. It is easier for criminals to only get a password than get both a password and a token response. Therefore, a multi-factor authentication gives more security, especially for ensuring that the transaction is done by the right person. However, an expert warned that the current malware has a way to circumvent it, for example, by infecting the victims' cellphones or making another phishing page in order to ask the victims to enter the authentication code they receive and then making a fraud transaction in background. According to another expert, there is a need to add something more besides the authentication factor, for example, digital signature as a mean of signing the transaction. Of course, there will be a trade-off between security and the convenience of customers.

An expert also highlighted that the uniqueness of the authentication mechanism of a bank compared to that of similar banks in the area could also bring an effect. For example, if there is a bank offering SMS authentication in a country where most of the banks offer token authentication, this bank might be ignored by the criminals as they focus more on developing a version of malware that attacks entities with token authentication.

Experts agree that domain popularity could be a factor in the target selection. They argued that criminals need to invest in malware that are capable of interacting with the dialogue of the bank. Therefore, targeting a large bank with many customers, which might be indicated from the popularity of its domain, is helpful as the malware they prepare could reach more potential victims, giving a better return for their investment. They referred to the top 3 banks in the Netherlands as an example. Meanwhile, one expert also provided another case of another Dutch bank Van Lanschot. The bank, as he explained, is intended for high-net worth individuals, but it only accounts for 2% of the market so it was not targeted by criminals. However, the attack is seen to be more spread out. Meanwhile, they thought that the domain popularity or the size of bank is only the case for phishing attacks and not for other means of social engineering. For example, a social engineering by phone does not consider whether the bank is large or not, or if its domain is more popular or not.

7.3.2. Experts' Opinion about the Models and Results

In general, experts found that the models look reasonable. Moreover, experts saw that this kind of model could be useful as a supplement to the risk model. However, they highlighted the limited number of factors included in the model and expressed that it is a bit dangerous to make a conclusion from the outcome because it did not look at the full scope of the attack since the model only considered the language, the authentication factor and some control variables. Experts believed that many other factors are more important than these factors in terms of explaining the target selection. For example, an expert highlighted the relative market share of the bank as a potential important factor. He argued that if the relative market share is compared together with the language factor, for example, it will diminish the significance of the language in explaining the target selection. Therefore, the conclusion drawn from the model generated without considering factors other than the language, authentication factor and the domain popularity, especially without considering any financial and market factor, could be misleading.

Security control might also influence the result. An expert used the "French model" to illustrate this issue. France used weak authentication in the past. However, at that time, the customers could only transfer money to a certain number of accounts. Moreover, adding an additional account could only be done manually. It was a different type of control but at that time it was very effective as one did not see a fraud. This model was adapted by Dutch banks by limiting the transfer amount of an account to another account which has no previous transactions record.

Besides, an expert shared his opinion about some language variables which were seen to have a negative relationship with the attack count in the model. He argued that the languages presented there link to countries that, according to him, seem to have relatively low density rate of online banking environment. It means that the society in those countries is not accustomed to online banking or not many banks there offer an online banking service. He implied that the language is not a single factor in this case. Other factors, like the maturity of online banking system of the country and the adoption rate of online banking by the society, also affect the relationship.

Another expert argued that some malware configurations are not used to attack the customers, but instead, to perform some analytics in order to determine whether they should invest in developing a web inject for the potential target or skip that target. Therefore, he suggested for the next research that the timestamp of the malware could be analysed as well.

In short, experts suggested that there were supposed to be more factors which might explain the target selection and, only by considering all possible factors, one can draw a plausible conclusion about the target selection. Also, more robust data is needed in order to make a better model.

7.4. Reflection on the Interview Results

Experts mentioned some other factors besides language and authentication factors. Some of them aligns with the characteristics identified in the literature review. For example, the record of successful attack attempts, once mentioned by an expert, is in line with the banks' attack records in the current list of characteristics. However, some of the factors could enhance the currently identified characteristics. For example, the presence of the relative market share of the bank characteristic can make us realise that the bank size characteristic can be represented by several more precise (sub-)characteristics. Meanwhile, there are factors that were not present in the list before. For instance, the expert mentioned the maturity of an online banking system of the country and the adoption rate of online banking by the society. However, it should be noted that these kinds of factors are made from the country-level perspectives, not the bank-level perspectives. Therefore, although they can be taken as factors that might affect target selection, they are beyond the characteristics that are attached to a bank. Knowing from which perspective the factors are from is important for researchers in order to approach the analysis from the correct point of view.

It is acknowledged that there are definitely more data and factors that could be added to the model which will consequently improve the model and the analysis. Furthermore, the better the model, the more plausible conclusion that can be made from it. However, the limitation with time and the availability of data makes it hard for researchers to come up with the most ideal model to explain the target selection. This suggests that the research of this topic should be iterative, which means that follow up studies should exist on top of this thesis so that better models and more acceptable conclusions about the target selection would be able to be drawn in the future.

8 Discussions and Recommendations

8.1. Discussion and Conclusion

Up to this point, lots of theoretical, quantitative and qualitative work have been performed in order to answer the research question. This section is going to revisit the research questions which were presented in the very beginning of the thesis.

The main research question of this thesis is:

What characteristics related to online banking services can affect the likelihood of the malware attack?

This thesis approached this main research question by conducting a quantitative research. This leads to the formulation of 4 research questions to help addressing the main research question. The argument for each research question will be given below.

RQ1. What characteristics of banks of their online banking services can potentially explain the likelihood of them to be targeted?

This research followed the below steps and processes to address RQ1:

- Literature study about the concept of cyberspace and cybercrime, the background, mechanism and security of online banking, the banking malware and actors in the financial malware attack.
- Literature study about the criminology theory, including the argument why the Routine Active Theory (RAT) is still relevant to the cybersecurity cases.
- Literature review in order to find the characteristics which potentially influence target selection of the banking malware, in accordance with relevant aspects of RAT.

The research identified several characteristics which potentially affect the target selection of the bank or its online banking service by the financial malware, which were listed in section 2.6. More factors were highlighted by experts from the interview, which could improve the list. Experts believed that the relative market share of banks could play a factor as they argued that the bank has a higher chance to be targeted if it has a higher relative market share. In addition, the maturity of online banking system of a country, the adoption rate of online banking by the society in a country, the number of bank registered in a country, banks' attack record and the quality of security control / quality of the detection system implemented by a bank are also mentioned. Some of these characteristics can be linked to one of the above characteristics. The relative market share of banks, for example, can be used to justify the size of banks and therefore can be added under the bank size characteristic. The maturity of online banking system, the adoption rate and the number of registered bank can be used to highlight the possible influence of the country where the bank is located to the target selection. Consequently, the quality of security control or the detection system of a bank can be incorporated with the rate of use of firewall/antivirus products as a group of characteristic named security control.

On the other hand, some arguments by the experts supported several findings from the literature. Experts seem to agree that a bank's attack record is one of potential characteristics as they argued that if a bank has been attacked successfully once, criminals will find it more attractive to be attacked in the future hoping that they will follow the same success. Meanwhile, experts are also seen to be in line with Kalige et al. (2012) about the argument that certain malware variants nowadays are able to circumvent the 2-factor authentication method.

Therefore, the list from the literature can be updated with inputs from the experts:

- **Related to value aspect:** bank size (in terms of number of customers, number of online users, total assets, total payments, revenues, net profit and relative market share), country location (in terms of financial status, number of Internet users, rate of banking/shopping penetration, the maturity of online banking system, the adoption rate of online banking in the country, the degree of cooperation between financial institutions and law enforcement, the degree of cooperation between financial institutions, money transfer policies of the country, the number of banks in the country and the availability of money mules within the country)

- **Related to visibility aspect:** brand popularity, domain name visibility, banks' attack record, website domain popularity, ownership of the bank, language of the online banking

- **Related to accessibility aspect:** bank authentication method, broadband penetration rate, users' online awareness, security control (in terms of rate of use of firewall/antivirus products and quality of detection system), ease in securely performing criminal actions

RQ2. What quantifiable metrics and variables relevant to the characteristics can be collected for the analysis?

Some characteristics were selected for the next step of the research, namely the language of the online banking, the authentication method and the website domain popularity. The latter was added for the control variable in the statistical models. In order to answer RQ2, data should be prepared beforehand. This research theoretically followed several processes to prepare the data, although in practice these processes were not executed straightforwardly due to some complications. The processes are:

- Analysing Fox IT database to understand the table relationships in order to extract the attack target data.

- Identifying and filtering the extracted dataset so that it includes the attack entries which correspond to banks in EU.

- Merging the attack target data with the list of EU banks so that the analysis could be performed at a bank level.

- Finding the external data such as language, authentication factor and domain popularity. Such data were obtained either by manual observation or from a legitimate source. An imputation technique was performed to missing values which might significantly reduce the quality of analysis.

The metrics and variables were extracted out of one or more of the data which were collected and processed. These are attack count, number of unique URLs corresponding to the bank, number of different malware variants targeting an online banking, the presence of a language in an online banking entity, the number of languages offered in an online banking entity, the presence of an authentication method in an online banking entity, and the popularity ranking or score of the domain.

This research also proposes the metric "week interval attack count" and "Unique attack ID count" which could be used to improve the traditional raw attack count in the context of approaching the actual attack count. This thesis also sees that the latter metric could potentially approach the actual attack count the best compared to the other proposed metrics since it is based on the way the malware attack data is stored rather than an unclarified assumption.

It is also seen from the cumulative distribution of all attack count metrics that, similar to a finding by Tajalizadehkhoob (Tajalizadehkhoob, 2013), the pareto power law still applies: a small proportion of banks (20%) accounts for more than 80% of attacks.

RQ3. How does the target landscape of online banks look like?

Descriptive analysis of the data provided nice insights for deciding on the possible and interesting metrics and variables to extract. First of all, the data is able to display the rise and fall of attacks performed by a malware. By using attack ID count, it is possible to reveal the non-overcounted attack trend of a malware.

It is also possible to see the distribution of malware attacks throughout countries in the EU. It is found that some malware variants are used to perform a specific attack targeting a specific bank or banks in a specific country. It is also found from the analysis on the period of attack that, besides having a specific target, some malware variants like ReactorBot and PkyBot performed attacks at a very specific time, indicating that they were specially designed to perform a personal attack.

Analysis on the proportion of the targeted and non-targeted online banking entities were also performed. From this analysis, it is seen that the proportion varied for every country. Bulgaria, for example, had nearly all of its banks targeted by the malware, while only a small proportion of banks in Ireland was targeted. It is also found that there is a clearly higher proportion of popular online bank domain in the targeted group than in the non-targeted one. In a similar way, the proportion of the presence of English and 2-factor authentication is also higher in the targeted group, even though the difference is not big.

RQ4. Which of the selected characteristics could explain the extent the online banking were targeted by the malware

Explanatory analysis was conducted in order to find out which of the extracted variables have a significant relationship with the target selection indicated by the attack count. A logistic regression model and negative binomial regression models were generated for this purpose.

In general, it is seen that some language variables are still significant despite the presence of control variables like the country, the domain popularity, the threat / malware variants and the number of unique attack URLs. For the languages, the variables are mostly not significant in the logistic regression model, indicating that there is no certainty that such variables can explain whether an online bank is targeted or not. However, provided that many language variables are significant in the negative binomial models, it can be inferred that some languages are able to explain whether an online banking will be more or less targeted.

It is also seen from the model that the English language is significant in explaining the target selection. The model result suggested that the presence of English increases the tendency of an online banking entity to be targeted. This finding is in line with the previous finding by Tajalizadehkhoob (2013), which showed that the presence of English was significant that time.

Meanwhile, the presence of two-factor authentication is also significant regardless of the control variables considered in the model. The negative coefficient in the negative binomial model indicates that the use of 2-factor authentication reduces the tendency of an online banking entity to receive more attacks. It may relate to the fact that 2-factor authentication provide more security barrier to the criminals than the 1-factor authentication. However, there is a significant, positive coefficient of this variable in the logistic model, meaning that there are more online banking entities applying 2-factor authentication that are targeted than those that not. The author argued that the criminals do not intentionally look for online banking entities with 2-factor authentication. They may search for big banks as their target and apparently many big banks nowadays implement 2-factor authentication. Therefore, it is appeared in the model like if the criminals attempt to target more entities with 2-factor authentication.

Expert interviews were also conducted. The interviews were intended to obtain experts' opinion on the target selection as well as to obtain their comments and interpretations of the models. In general, experts think that although the characteristics that were assessed in this research may be able to explain the target selection, it is too premature to form that conclusion since the argument was extracted from a limited model. Experts argued that in order to come up with plausible conclusions about the target selection, the model should be improved by adding other factors, which means that more data should be added. Experts believed that other factors, especially financial- and market-related factors like the relative market share of the bank, are more influential than language and authentication method. Despite this limitation, they find the result pretty rational.

8.1.1. Conclusion

With respect to the main research question, this thesis concludes that:

> There are several characteristics of banks' online banking services that potentially can explain the likelihood of them being attacked by the banking malware. Some of these characteristics were assessed, and it was found that these characteristics, for example the presence of English and 2-factor authentication, were seen to be adequately significant in affecting the likelihood of the attack, despite the presence of the control variables in the model. However, in order to make a plausible conclusion about their influence, according to the experts, improving the current statistical models by considering more factors, especially financial- and market-related ones, is required.

8.2. Scientific and Social Contribution of this Thesis

There are several scientific contributions that this thesis gives, especially for the continuation of the research related to this topic in TU Delft:

- The attack target dataset used by this thesis is broader than the one used in previous related thesis in TU Delft. It not only comprises ZeuS, but also other malware variants. Moreover, The time frame of the dataset, which is from February 2014 to November 2017, is relatively newer than the previous dataset. Therefore, this thesis could give more updated insights about the target selection compared to previous studies.

- This thesis introduces approaches of extracting the attack target data from the database, filter the dataset and process it in order to make it useful for the analysis. This approaches may hopefully help the next researchers in their analysis when using the same data.

- Unlike other theses which only look at the targeted entities, this thesis also puts the non-targeted EU banks into account.

- This thesis also introduces other metrics to count malware attacks. One of them is the attack ID count metric, which is believed to approach the counting of actual attacks much better.

- This thesis is also among the first to evaluate the relationship between language, authentication factor and the attack count which serves as an indicator for explaining the target selection by this kind of malware attack.

This thesis might be suitable as a stepping stone for those who want to do research in the similar topic and in the similar fashion. There are still lots more characteristics that can be assessed, which in the end could enhance the result of this thesis. This thesis also presents issues, complications, difficulties and pitfalls encountered during the execution of the research. Hopefully, the documentation of such elements could help the next researchers solve or avoid them in the future.

Meanwhile, this thesis may also contribute to the society, in particular the banks in EU and the security companies like Fox IT. For the banks in EU, this thesis might be useful in order to recalibrate their position in the malware threat landscape so that they become more aware of the potential of the attack towards them. For the security companies, insights that were obtained from this thesis can help them to understand more about the threat landscape and the target selection. For instance, the security companies can now know whether a malware variant was used to perform generic attacks or a specialized attack, and if the malware variant performed a specialized attack, which country the malware variants targeted. This kind of insight can be useful for them to determine which area should they put more attention into in terms of eradicating the malware attack. Moreover, similar to the scientific contribution of this thesis, the security companies can also use this thesis as a basis in developing better analysis for understanding the target selection. In particular, they can enhance the statistical model in this thesis so that more acceptable conclusion about which characteristics influence the target selection can be obtained in the future. The conclusion from a better model will help the security companies in giving reliable recommendations to their clients.

Based on the limited analysis and model, this thesis tries to provide a recommendation regarding the implementation of two-factor authentication. The model result suggests that the use of 2-factor authentication reduces the tendency of an online banking entity to receive more attacks. It is also inferred from the interview that experts still favour the application of two-factor authentication as it provides more sense of security. Therefore, banks are suggested to implement or maintain a 2-factor authentication in their online banking services.

8.3. Limitation and Future Research

Limitations in this thesis have been mentioned during the explanation of phases and steps in previous chapters. This section is going to summarize those limitations and provide arguments about the implication they are bringing to the result of this thesis. The following points specify and explain about the limitations that are present in this thesis:

- Due to the difficulty in collecting relevant data, especially financial and market data for all EU banks as in the list, and also the interpretability issues if all possible factors are included in the model, variables included in the model for this thesis are limited to the attack-related characteristics (e.g. unique URLs, malware variants), language, authentication factor and some control variables. Thus, this model does not take any financial factor into account. The experts argued that financial and market factors are the most determining factors in the target selection of this type of malware attack and hence it would be dangerous to conclude the target selection only from the result of this model.

  The author also realised that this model alone cannot be used to explain the target selection completely. In order to explain it more comprehensively, it needs to be paired with other studies that research other characteristics of the online banking. The future research may want to compare the outcome of this thesis with other studies in the target selection topic.

Besides, given that the data can be collected, adding financial and market factors to the model can be a step to improve this thesis in the future.

- This thesis also highlights a limitation in terms of data quality. It was explained that this thesis relies on data and entity information that came from the entity identification process performed by Fox IT. It was also found that the identification process itself is not perfect since some false negatives are present. The inability of the mechanism to identify all attack URLs that correspond to bank entities can imply that the result which is based on such data might be not perfect as well. The effect of this limitation was reduced by applying domain extraction and matching mechanisms. However, it cannot completely remove the false negative issue in the dataset.

  Good analysis relies on good data. Therefore, it is recommended that there is an initiative in the future to design and build better mechanisms for identifying bank domains and entities from URL, so that the false negative can be reduced. Subsequently, such study can be done again using better data, which hopefully will result in better results.

- Some processes in this thesis were performed manually. As in other manual activities, human factor could play a role. There is a potential human error in the process, although it is hard to calculate. Human error in the data collection phase could reduce the quality of the data collected which might affect the quality of the analysis result. Considering that many data to explain the target selection are difficult to obtain, collecting the data manually might be the only feasible way with some cases.

  Groups of researchers in this field are recommended to invest in getting an access to a reliable data source or making a reusable repository to collect and store the observed data. Such investment will be useful for future researchers to make a better analysis without getting too much hassle on the availability and reliability of the data. The researchers should also make sure that they have the data they need to examine the characteristics or factors which they focus on in advance.

- It was described in the chapter about data preparation for model generation that there were several data points which were removed due to the lack of language and authentication factor information because of no online banking service for the banks. This might also affect the model as well as the result.

- Limitation on the metrics: advantages and limitations were given previously to each metric generated for approaching the actual attack count. Special attention is given to the metric week-interval attack count. It is heavily based on the assumption that the interval for criminals to make an exactly the same actual attack (the same attack URLs, the same malware variants) is 7 days. It was developed to deal with the problem of overestimation as in the raw attack count. However, the assumption used for this metric has been questioned, with the argument that many things should be checked in order to ensure that the assumption can be applied or not.

This thesis also proposes several potential future studies, which are extracted from the above limitations as well as any input given for this thesis:

- Enhance this thesis by adding financial and market factors to the model. Especially, experts mentioned a factor "relative market share of the bank" as another important factor that might explain the target selection. It will be valuable if the future research could focus on assessing this factor. Besides, factors like "the number of financial institutions in the country", "the maturity of online banking system of a country", "the adoption rate of online banking in a country" and "the quality of banks' detection system", as also highlighted by the experts, could also be interesting to research.

- Extend this thesis with other studies that research other factors of the online banking to make a more comprehensive result so that a better conclusion can be drawn. The extension is suggested to not only limited to incorporating the results, but also the integration of multiple data and sharing of the processes. This thesis recommends the creation of repositories to store various data that can be easily used for the future research on this topic.
- Conduct a study to design and build improved mechanisms or algorithms for identifying bank domains and entities out of the attack URLs.
- Execute the same quantitative analysis using improved data source in order to obtain a better model result and more reliable analysis.

References

Anderson, R., Barton, C., Böhme, R., Clayton, R., van Eeten, M. J. G., Levi, M., … Savage, S. (2013). Measuring the Cost of Cybercrime. In R. Böhme (Ed.), *The economics of information security and privacy* (pp. 265–300). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-39498-0_12

Andriesse, D., Rossow, C., Stone-Gross, B., Plohmann, D., & Bos, H. (2013). Highly resilient peer-to-peer botnets are here: An analysis of Gameover Zeus. *Proceedings of the 2013 8th International Conference on Malicious and Unwanted Software: "The Americas", MALWARE 2013*, 116–123. doi: 10.1109/MALWARE.2013.6703693

Asghari, H. (2010). *Botnet mitigation and the role of ISPs: A quantitative study into the role and incentives of Internet Service Providers in combating botnet propagation and activity* (Master's Thesis). Delft University of Technology.

Asghari, H., van Eeten, M., & Bauer, J. M. (2016). Economics of cybersecurity. In *Handbook on the economics of the internet* (p. 262-287). Edward Elgar Publishing.

Bach, O. (2015). *Tinba: World's Smallest Malware Has Big Bag of Nasty Tricks.* Retrieved 2018-05-30, from `https://securityintelligence.com/tinba-worlds-smallest-malware-has-big-bag-of-nasty-tricks/`

Banks, E. (2001). *E-Finance: The Electronic Revolution in Financial Services.* New York, NY, USA: John Wiley & Sons, Inc.

Beazley Breach. (2016). *Hackers target smaller financial institutions* (Tech. Rep.). Beazley. Retrieved 2018-03-25, from `https://www.beazley.com/documents/Insights/201607-hackers-target-smaller-financial-institutions.pdf`

Berghoff, T. (2017). *Analysis: ZeuS Panda.* Retrieved 2018-03-11, from `https://www.gdatasoftware.com/blog/2017/08/29928-analysis-zeus-panda`

Bisson, D. (2017). *How TrickBot Malware's Code and Delivery Methods Evolved in Q3 2017.* Retrieved 2018-07-19, from `https://www.tripwire.com/state-of-security/latest-security-news/trickbot-malwares-code-delivery-methods-evolved-q3-2017/`

Bottazzi, G., & Me, G. (2014). The Botnet Revenue Model. In *Proceedings of the 7th international conference on security of information and networks* (pp. 459:459—-459:465). New York, NY, USA: ACM. doi: 10.1145/2659651.2659673

Bougaardt, G., & Kyobe, M. (2011). Investigating the factors inhibiting SMEs from recognizing and measuring losses from cyber crime in South Africa. In *Icime 2011-proceedings of the 2nd international conference on information management and evaluation: Icime 2011 ryerson university, toronto, canada, 27-28 april 2011* (p. 62).

Boutin, J.-I. (2013). *Qadars – a banking Trojan with the Netherlands in its sights.* Retrieved 2018-05-30, from `https://www.welivesecurity.com/2013/12/18/qadars-a-banking-trojan-with-the-netherlands-in-its-sights/`

Brenner, S. W. (2006). At light speed: Attribution and response to cybercrime/terrorism/warfare. *J. Crim. L. & Criminology*, *97*, 379.

Calisir, F., & Gumussoy, C. A. (2008). Internet banking versus other banking channels: Young consumers' view. *International Journal of Information Management*, *28*(3), 215–221. doi: 10.1016/j.ijinfomgt.2008.02.009

Campanella, F., Della Peruta, M. R., & Del Giudice, M. (2017, mar). The Effects of Technological Innovation on the Banking Sector. *Journal of the Knowledge Economy*, *8*(1), 356–368. Retrieved from `https://doi.org/10.1007/s13132-015-0326-8` doi: 10.1007/s13132-015-0326-8

Capeller, W. (2001). Not such a neat net: Some comments on virtual criminality. *Social and Legal Studies*, *10*(2), 229–242. doi: 10.1177/a017404

Carman, A. (2015). *Dyre malware infections surge in 2015.* Retrieved 2018-07-19, from `https://www.scmagazine.com/trend-micro-documents-new-malware-infections/article/534014/`

Cheung, R. (2017). *Targeting financial organisations: a multi-sided perspective* (Master's Thesis). Delft University of Technology.

Cimpanu, C. (2016). *Qadars Trojan Returns Bigger and Badder than Ever Before.* Retrieved 2018-05-30, from `https://news.softpedia.com/news/qadars-trojan-returns-bigger-and-badder-than-ever-before-508546.shtml`

Cisco Umbrella. (n.d.). *Cisco Popularity List.* Retrieved 2018-06-13, from `http://s3-us-west-1.amazonaws.com/umbrella-static/index.html`

Claessens, J., Dem, V., De Cock, D., Preneel, B., & Vandewalle, J. (2002, jun). On the Security of Today's Online Electronic Banking Systems. *Computers & Security*, *21*(3), 253–265. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0167404802003127http://linkinghub.elsevier.com/retrieve/pii/S0167404802003127` doi: 10.1016/S0167-4048(02)00312-7

Clarke, R. V. G., & Felson, M. (1993). *Routine activity and rational choice: Advances in criminological theory* (Vol. 5). Piscataway, NJ. Retrieved from `http://discovery.lib.harvard.edu/?itemid={%}7Clibrary/m/aleph{%}7C012555038`

Cohen, L. E., & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, *44*(4), 588. Retrieved from `http://www.jstor.org/stable/2094589?origin=crossref` doi: 10.2307/2094589

Craig, D. (2016). *Five Cybersecurity Challenges Facing Financial Services Organizations Today.* Retrieved 2017-11-03, from `https://securityintelligence.com/five-cybersecurity-challenges-facing-financial-services-organizations-today/`

Creswel, J. W. (2008). The Selection of a Research Approach. *Research design: qualitative, quantitative, and mixed methods approaches*, 3–22. doi: 45593:01

Crisanto, J. C., & Prenio, J. (2017). Regulatory approaches to enhance banks' cyber-security frameworks. *Bank for International Settlements*(2).

Cucu, P. (2017). *How A Banking Trojan Does More Than Just Steal Your Money.* Retrieved 2018-03-25, from `https://heimdalsecurity.com/blog/banking-trojan/`

D'Alfonso, S. (2014). *Who Are 'Knowing' Money Mules?* Retrieved 2018-03-08, from https://securityintelligence.com/who-are-knowing-money-mules/

Ducklin, P. (2016). *Gozi virus author finally sentenced – should be out and home soon – Naked Security.* Retrieved 2018-05-30, from https://nakedsecurity.sophos.com/2016/01/07/gozi-virus-author-finally-sentenced-should-be-out-and-home-soon/

Etaher, N., Weir, G. R., & Alazab, M. (2015). From ZeuS to zitmo: Trends in banking malware. *Proceedings - 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2015, 1*, 1386–1391. doi: 10.1109/Trustcom.2015.535

European Banking Authority. (2014). *Final Guidelines on the Security of Online Payments* (Tech. Rep. No. December).

European Central Bank. (2018). *MFI statistical report.* Retrieved 2018-03-21, from https://www.ecb.europa.eu/stats/financial_corporations/list_of_financial_institutions/html/daily_list-MID.en.html

Ezeoha, A. E. (2006). Regulating Internet Banking in Nigeria: some success prescriptions–part 2. *Journal of Internet Banking and Commerce, 11*(1), 35–47.

Falliere, N., & Chien, E. (2009). *Zeus: King of the Bots* (Tech. Rep.). Symantec Security Response. Retrieved from http://courses.isi.jhu.edu/malware/papers/ZEUS.pdf

Federal Bureau of Investigation. (2010). *FBI — Cyber Bust.* Retrieved 2018-03-08, from https://archives.fbi.gov/archives/news/stories/2010/october/cyber-banking-fraud

Fox-IT. (n.d.). *Fox-IT | For a more secure society.* Retrieved 2018-03-27, from https://www.fox-it.com/en/

Fox-IT. (2013). *Analysis of the KINS malware | Fox-IT International blog.* Retrieved 2018-05-30, from https://blog.fox-it.com/2013/07/25/analysis-of-the-kins-malware/

Gelman, A., & Hill, J. (2006). *Missing-data imputation.* Cambridge: Cambridge University Press. Retrieved from http://ebooks.cambridge.org/ref/id/CBO9780511790942A231 doi: 10.1017/CBO9780511790942.031

Gilleland, M. (2006). *Levenshtein Distance.* Retrieved 2018-05-30, from https://people.cs.pitt.edu/{~}kirk/cs1501/assignments/editdistance/Levenshtein%20Distance.htm

Gomber, P., Koch, J.-A., & Siering, M. (2017). Digital Finance and FinTech: current research and future research directions. *Journal of Business Economics, 87*(5), 537–580. Retrieved from http://link.springer.com/10.1007/s11573-017-0852-x doi: 10.1007/s11573-017-0852-x

Gopalakrishnan, S., Wischnevsky, J. D., & Damanpour, F. (2003). A multilevel analysis of factors influencing the adoption of Internet banking. *IEEE Transactions on Engineering Management, 50*(4), 413–426.

Gordon, S., & Ford, R. (2006). On the definition and classification of cybercrime. *Journal in Computer Virology, 2*(1), 13–20.

Grabosky, P. M. (2001). Virtual Criminality: Old Wine in New Bottles? *Social & Legal Studies, 10*(200106), 243–249.

Hutchings, A., & Clayton, R. (2017). Configuring Zeus: A case study of online crime target selection and knowledge transmission. *eCrime Researchers Summit, eCrime*, 33–40. doi: 10.1109/ECRIME.2017.7945052

Hutchinson, D., & Warren, M. (2003). Security for Internet banking: a framework. *Logistics Information Management*, *16*(1), 64–73. Retrieved from `http://www.emeraldinsight.com/doi/10.1108/09576050310453750` doi: 10.1108/09576050310453750

Infosecurity Magazine. (2010). *Zeus is king of bank fraud trojan viruses - Infosecurity Magazine.* Retrieved 2018-03-08, from `https://www.infosecurity-magazine.com/news/zeus-is-king-of-bank-fraud-trojan-viruses/`

IZOOlogic. (2016). *Top 4 Malware – Financial Trojans – Zeus, Carberp, Citadel and SpyEye.* Retrieved 2018-02-28, from `https://www.izoologic.com/2016/10/15/top-4-malware-financial-trojans-zeus-carberp-citadel-and-spyeye/`

Jaleshgari, R. P. (1999). Document trading online. *Information Week*, *755*(136), 136.

Kalige, E., Burkey, D., & Director, I. P. S. (2012). A case study of Eurograbber: How 36˜{m}illion euros was stolen via malware. *Versafe (White paper)*(December).

Kaspersky. (2018). *Financial Cyberthreats in 2017* (Tech. Rep.). Author. Retrieved from `https://d2538mqrb7brka.cloudfront.net/wp-content/uploads/sites/43/2018/03/07162608/Kaspersky_Lab_financial_cyberthreats_in_2017.pdf`

Kaspersky Lab. (2017). *The Dridex Banking Trojan: an ever-evolving threat | Kaspersky Lab.* Retrieved 2018-04-11, from `https://www.kaspersky.com/about/press-releases/2017{_}the-dridex-banking-trojan-an-ever-evolving-threat`

Kiljan, S., Simoens, K., De Cock, D., van Eekelen, M., & Vranken, H. (2014). Security of Online Banking Systems. *Tech. Rep. TR-OU-INF-2014-01 (Open Universiteit).*

Kiljan, S., Vranken, H., & van Eekelen, M. (2018). Evaluation of transaction authentication methods for online banking. *Future Generation Computer Systems*, *80*, 430–447. Retrieved from `http://dx.doi.org/10.1016/j.future.2016.05.024` doi: 10.1016/j.future.2016.05.024

Kiljan, S. Z. (2017). *Exploring, Expanding and Evaluating Usable Security in Online Banking.* Open Universiteit.

Lagazio, M., Sherif, N., & Cushman, M. (2014). A multi-level approach to understanding the impact of cyber crime on the financial sector. *Computers and Security*, *45*(0), 58–74. doi: 10.1016/j.cose.2014.05.006

Leukfeldt, E. R., & Yar, M. (2016). Applying Routine Activity Theory to Cybercrime: A Theoretical and Empirical Analysis. *Deviant Behavior*, *37*(3), 263–280. doi: 10.1080/01639625.2015.1012409

Lowe, G. (1997, jun). A hierarchy of authentication specifications. In *Proceedings 10th computer security foundations workshop* (pp. 31–43). doi: 10.1109/CSFW.1997.596782

Lu, Z., Marvel, L., & Wang, C. (2015). To be proactive or not: a framework to model cyber maneuvers for critical path protection in MANETs. In *Proceedings of the second acm workshop on moving target defense* (pp. 85–93).

Malwarebytes Labs. (2017a). *Inside the Kronos malware - part 1.* Retrieved 2018-04-11, from `https://blog.malwarebytes.com/cybercrime/2017/08/inside-kronos-malware/`

Malwarebytes Labs. (2017b). *Inside the Kronos malware - part 2.* Retrieved 2018-04-11, from `https://blog.malwarebytes.com/cybercrime/2017/08/inside-kronos-malware-p2/`

Mannan, M., & van Oorschot, P. C. (2008). Security and usability: the gap in real-world online banking. In *Proceedings of the 2007 workshop on new security paradigms* (pp. 1–14).

Nagurney, A. (2015). A multiproduct network economic model of cybercrime in financial services. *Service Science, 7*(1), 70–81.

Nand, P., Astya, R., & Singh, D. (2015). An Add-on to Present Banking: m-banking. In *Proceedings of fourth international conference on soft computing for problem solving , advances in intelligent systems and computing* (Vol. 336, pp. 377–388). Springer India. Retrieved from `http://link.springer.com/10.1007/978-81-322-2220-0` doi: 10.1007/978-81-322-2220-0

OECD. (2008). Malicious Software (Malware):A security Threat to the Internet Economy. *OECD*, 1–106.

OpenRefine. (2018). *OpenRefine.* Retrieved 2018-05-30, from `http://openrefine.org/`

Petee, T. A., Corzine, J., Huff-Corzine, L., Clifford, J., & Weaver, G. (2010). Defining "cyber-crime": Issues in determining the nature and scope of computer-related offenses. *Futures Working Group, 5*, 6–11.

PlugandPlay Tech Center. (2017). *The Cybersecurity Threats Facing Financial Institutions.* Retrieved from `http://plugandplaytechcenter.com/2017/06/26/cybersecurity-threats-financial-institutions/`

PwC. (2018). *Pulling fraud out of the shadows: Global Economic Crime and Fraud Survey 2018* (Tech. Rep.). PricewaterhouseCoopers.

Raghavan, A. R., & Parthiban, L. (2014). The effect of cybercrime on a bank's finances. *International Journal of Current Research and Academic Review, 2*(2), 173–178.

Sanghavi, M. (2015). *DRIDEX and how to overcome it.* Retrieved 2018-04-11, from `https://www.symantec.com/connect/blogs/dridex-and-how-overcome-it`

Schwarz, D. (2015). *Peeking at Pkybot.* Retrieved 2018-05-26, from `https://asert.arbornetworks.com/peeking-at-pkybot/`

Secureworks Counter Threat Unit Threat Intelligence. (2016). *Banking Botnets: The Battle Continues | Secureworks.* Retrieved 2018-05-26, from `https://www.secureworks.com/research/banking-botnets-the-battle-continues`

Sherstobitoff, R. (2012). Inside the world of the citadel trojan. *Emergence, 9*.

Singer, P. W., & Friedman, A. (2014). *Cybersecurity: What everyone needs to know.* Oxford: Oxford University Press.

Symantec. (n.d.). *W32.Ramnit.* Retrieved 2018-04-11, from `https://www.symantec.com/security-center/writeup/2010-011922-2056-99`

Tajalizadehkhoob, S. (2013). *Online Banking Fraud Mitigation: A Quantitative Study for Extracting Intelligence about Target Selection by Cybercriminals from Zeus Financial Malware Files* (Master's Thesis). Delft University of Technology.

Tajalizadehkhoob, S., Asghari, H., Gañán, C., & van Eeten, M. (2014). Why Them? Extracting Intelligence about Target Selection from Zeus Financial Malware. *Workshop on the Economics of Information Security (WEIS)*, 1–26.

Tajalizadehkhoob, S., Böhme, R., Gañán, C., Korczynski, M., & van Eeten, M. (2017). Rotten apples or bad harvest? what we are measuring when we are measuring abuse. *CoRR*, *abs/1702.01624*. Retrieved from `http://arxiv.org/abs/1702.01624`

Tajalizadehkhoob, S., Gañán, C., Noroozian, A., & van Eeten, M. (2017). The Role of Hosting Providers in Fighting Command and Control Infrastructure of Financial Malware. In *Proceedings of the 2017 acm on asia conference on computer and communications security* (pp. 575–586). New York, NY, USA: ACM. doi: 10.1145/3052973.3053023

Thomas, D., & Loader, B. D. (2000). *Introduction - cybercrime: law enforcement, security and surveillance in the information age, Cybercrime: Law Enforcement, Security and Surveillance in the Information Age.* Taylor & Francis Group, New York, NY.

TrendMicro. (2012). *Security Threats to Business, the Digital Lifestyle, and the Cloud* (Tech. Rep.). Retrieved from `http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/spotlight-articles/sp-trendmicro-predictions-for-2013-and-beyond.pdf`

Van Moorsel, D. (2016). *Target selection regarding financial malware attacks within the Single Euro Payments Area* (Master's Thesis). Delft University of Technology.

van den Berg, J., van Zoggel, J., Snels, M., van Leeuwen, M., Boeke, S., van de Koppen, L., ... Bos, T. D. (2014). On (the Emergence of) Cyber Security Science and its Challenges for Cyber Security Education. *NATO STO/IST-122 symposium in Tallin*(c), 1–10.

Voolf, D., Boddy, S., & Smith, J. (2017). *Trickbot Focuses on Wealth Management Services from its Dyre Core.* Retrieved 2018-07-29, from `https://www.f5.com/labs/articles/threat-intelligence/trickbot-focuses-on-wealth-management-services-from-its-dyre-core`

Vrancianu, M., & Popa, L. A. (2010). Considerations Regarding the Security and Protection of E-Banking Services Consumers' Interests. *The AMFITEATRU ECONOMIC journal*, *12*(28), 388–403. Retrieved from `http://ideas.repec.org/a/aes/amfeco/v12y2010i28p388-403.html`

Wikström, P.-O. H. (2011). Does everything matter? Addressing the problem of causation and explanation in the study of crime. *When Crime Appears: The Role of Emergence*, 53–62. Retrieved from `http://www.routledge.com/books/details/9780415883054/`

Wikström, P.-O. H., & Treiber, K. (2016). Situational theory: The importance of interactions and action mechanisms in the explanation of crime. *The handbook of criminological theory*, 415–444.

Wueest, C. (2016). *Security Response: Financial threats 2015* (Tech. Rep.). Symantec. Retrieved from `https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/financial-threats-15-en.pdf`

Wyke, J. (2011). *What is Zeus?* (Tech. Rep.). SophosLab. Retrieved from `https://www.sophos.com/en-us/medialibrary/pdfs/technical%20papers/sophos%20what%20is%20zeus%20tp.pdf`

Wyke, J. (2012). *The Citadel crimeware kit – under the microscope – Naked Security.* Retrieved 2018-03-11, from `https://nakedsecurity.sophos.com/2012/12/05/the-citadel-crimeware-kit-under-the-microscope/`

Xu, Y., Bailey, M., Vander Weele, E., & Jahanian, F. (2010). CANVuS: Context-aware network vulnerability scanning. In *International workshop on recent advances in intrusion detection* (pp. 138–157).

Yar, M. (2005). The Novelty of 'Cybercrime': An Assessment in Light of Routine Activity Theory. *European Journal of Criminology*, *2*(4), 407–427.

A

Procedure to extract second-level domains (in Python)

```
## Functions to extract domains
from tldextract import extract as tldextract

#Function to remove the http, https of site_url
def strip_site(site):
    """Removes leading http:// or https:// and trailing ’/’"""
    site = site.lower()
    site = site[1:] if site.startswith(’\"’) else site #remove " in the beginning of url if any
    site = site[1:] if site.startswith(’#’) else site #remove # in the beginning of url if any
    site = site[1:] if site.startswith(’!’) else site #remove ! in the beginning of url if any
    site = site[1:] if site.startswith(’^’) else site #remove ! in the beginning of url if any
    site = site[:-1] if site.endswith(’\"’) else site #remove " in the end of url if any
    site = site[:-1] if site.endswith(’$’) else site #remove $ in the end of url if any

    #get rid of starting star (*). Taken from Samaneh’s thesis
    if site.startswith(’*.’):
        # e.g.: https://*.westpac.com.au/esis/login/srvpage*
        site = site[2:]
    elif site.startswith(’*’) and site[:2]!=’*/’:
        # Avoid stripping e.g. https://*/xxx ....
        # This can in some cases still cause problems, e.g.: *-sparkasse.de* (not handled yet)
        site = site[1:]

    #get rid of ending star (*). Taken from Samaneh’s thesis
    if len(site)>2 and site[-1]==’*’ and site[-2]!=’.’:
        # throw out trailing *
        site = site[:-1]

    site = site.replace('#.','.')
    if site.startswith('http'): #remove http*
        site = site.replace('https://', '').replace('http*://', '
           ').replace('http://', '').replace('http*//', '').
           replace('http:/', '').replace('https*', '').replace('
           http*', '')
    if "www??" or "www?" in site: #change www?? or www? to www
       because it cause tldextract to make identification mistake
        site = site.replace("www??","www").replace("www?","www")
    if site.endswith('/'):
        site = site[:-1]
    return site

#Function to convert star sign (*) to dot (.)
def convert_regex_sign(site):
    """Convert any regex-specific sign e.g.: \. \- \/ .* etc."""
    site = site.replace('(^|\.)','.')  #regex (^|\.) (start
       character or .) to .
    site = site.replace('(//|\.)','.') #regex (//|\.) (double
       slash or .) to .
    site = site.replace('(.*?)','*') #regex (.*?) (zero or more
       characters, non-greedy, captured) to *
    site = site.replace('(.*)','*') #regex (.*) (zero or more
       characters,captured) to *
    site = site.replace('.*?','*') #regex .*? (zero or more
       characters, non-greedy) to *
    site = site.replace('\.','.') #regex \. to .
    site = site.replace('\/','/') #regex \/ to /
    #site = site.replace('.*','*') #regex .* (zero or more
       characters) to *

    return site

#Function to strip and extract the domains of site_url
def extract_domain(site):
    """Returns domain+tld from a full domain"""
    site = site.strip() #remove whitespaces in the beginning and
       end of url
    site = convert_regex_sign(site) # apply convert_regex_sign
       function
    site = strip_site(site)  # just to be sure
    ext = tldextract(site)
    ret = ".".join([ss for ss in ext[-2:] if ss])
    # In case of gov.ie, or IPs, our 'ret' works better than 'ext
       .registered_domain'
    return ret
```

```

#Function to extract the domains of host
def extract_host(host):
    """Returns domain+tld from a full domain"""
    ext = tldextract(host)
    ret = ".".join([ss for ss in ext[-2:] if ss])
    return ret
```

B

Summary of Regression Model

B.1. Logistic Regression

```
Call:
glm(formula = is_targeted ~ langEnglish + langGerman + langFrench +
    langDutch + langItalian + langSpanish + langPortugese + langGreek +
    langCzech + langSlovak + langSlovenian + langPolish + langHungarian +
    langRomanian + langBulgarian + langDanish + langSwedish +
    langFinnish + langLatvian + langEstonian + langLithuanian +
    lang_count + auth1FA + auth2FA + pop_score + Country, family = "binomial",
    data = data_logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
−3.1156  −0.8775  −0.2468   1.1215   3.5663

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         −8.561e−01  4.860e−01  −1.762 0.078113 .
langEnglishTrue      6.286e−01  3.142e−01   2.001 0.045420 *
langGermanTrue       5.152e−01  3.851e−01   1.338 0.180985
langFrenchTrue       3.548e−01  4.551e−01   0.780 0.435605
langDutchTrue        6.035e−01  5.615e−01   1.075 0.282397
langItalianTrue      8.465e−01  6.395e−01   1.324 0.185564
langSpanishTrue     −6.052e−02  6.252e−01  −0.097 0.922880
langPortugeseTrue    8.850e−01  1.022e+00   0.866 0.386715
langGreekTrue        1.160e+00  1.211e+00   0.958 0.337857
langCzechTrue        9.680e−02  1.106e+00   0.087 0.930285
langSlovakTrue       2.075e+00  1.074e+00   1.932 0.053340 .
langSlovenianTrue   −1.987e+00  1.817e+00  −1.094 0.274011
langPolishTrue      −1.573e+00  1.623e+00  −0.970 0.332247
langHungarianTrue    1.987e+00  1.561e+00   1.273 0.203008
langRomanianTrue    −1.735e+01  3.917e+02  −0.044 0.964662
langBulgarianTrue    1.196e+01  1.485e+03   0.008 0.993572
langDanishTrue      −5.425e−01  1.197e+00  −0.453 0.650465
```

```
langSwedishTrue        2.146e+00  7.926e-01   2.708  0.006777 **
langFinnishTrue       -1.894e+00  1.335e+00  -1.419  0.155779
langLatvianTrue        2.482e-01  1.568e+00   0.158  0.874244
langEstonianTrue      -3.439e+00  9.157e-01  -3.756  0.000173 ***
langLithuanianTrue     1.597e+01  2.400e+03   0.007  0.994690
lang_count            -3.465e-02  2.579e-01  -0.134  0.893119
auth1FATrue            4.577e-01  2.861e-01   1.600  0.109644
auth2FATrue            2.148e+00  2.658e-01   8.084  6.26e-16 ***
pop_score              4.466e-06  2.956e-07  15.111   < 2e-16 ***
CountryBelgium        -1.844e+00  7.015e-01  -2.629  0.008574 **
CountryBulgaria        1.413e+01  1.370e+03   0.010  0.991771
CountryCroatia         7.971e-01  8.065e-01   0.988  0.322969
CountryCyprus         -1.879e+00  8.693e-01  -2.161  0.030681 *
CountryCzechia        -3.515e-01  1.204e+00  -0.292  0.770389
CountryDenmark        -4.040e+00  1.265e+00  -3.193  0.001407 **
CountryEstonia                NA         NA      NA        NA
CountryFinland         1.552e+00  1.358e+00   1.143  0.253238
CountryFrance         -6.775e-01  5.497e-01  -1.233  0.217755
CountryGermany        -9.497e-02  2.994e-01  -0.317  0.751092
CountryGreece         -2.406e+00  1.277e+00  -1.883  0.059636 .
CountryHungary        -5.971e+00  1.576e+00  -3.790  0.000151 ***
CountryIreland        -3.671e+00  5.120e-01  -7.171  7.45e-13 ***
CountryItaly          -8.979e-01  7.032e-01  -1.277  0.201631
CountryLatvia         -1.435e+00  1.623e+00  -0.884  0.376778
CountryLithuania      -1.819e+01  2.400e+03  -0.008  0.993952
CountryLuxembourg     -2.082e+00  5.290e-01  -3.936  8.27e-05 ***
CountryMalta          -3.517e+00  1.032e+00  -3.410  0.000650 ***
CountryNetherlands    -1.343e+00  6.935e-01  -1.937  0.052804 .
CountryPoland         -1.455e-01  1.632e+00  -0.089  0.928978
CountryPortugal       -6.810e+00  1.073e+00  -6.349  2.17e-10 ***
CountryRomania         1.688e+01  3.917e+02   0.043  0.965613
CountrySlovakia       -1.136e+00  1.035e+00  -1.098  0.272076
CountrySlovenia       -7.328e-01  2.003e+00  -0.366  0.714491
CountrySpain           3.515e-01  7.017e-01   0.501  0.616451
CountrySweden         -3.840e+00  8.665e-01  -4.431  9.37e-06 ***
CountryUnited Kingdom -1.060e+00  4.437e-01  -2.390  0.016847 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5319.3  on 3850  degrees of freedom
Residual deviance: 3691.7  on 3799  degrees of freedom
AIC: 3795.7

Number of Fisher Scoring iterations: 15
```

B.2. Negative Binomial Regression towards Raw Attack Count

```
Call:
glm.nb(formula = raw_attack_count ~ langEnglish + langGerman +
    langFrench + langDutch + langItalian + langSpanish + langPortugese +
    langGreek + langCzech + langSlovak + langSlovenian + langPolish +
    langHungarian + langRomanian + langBulgarian + langDanish +
    langSwedish + langFinnish + langLatvian + langEstonian +
    langLithuanian + lang_count + auth1FA + auth2FA + pop_score +
    Country + threat_name + year + unique_attackurl_count,
    data = data_notzero, init.theta = 1.033310472, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
−3.2969  −0.9894  −0.2593   0.1431   9.0820

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   −2.704e+00  2.842e−01  −9.513  < 2e−16 ***
langEnglishTrue                3.788e−01  2.923e−02  12.957  < 2e−16 ***
langGermanTrue                 2.651e−01  3.750e−02   7.070 1.55e−12 ***
langFrenchTrue                 9.960e−03  4.865e−02   0.205 0.837788
langDutchTrue                  2.771e−01  6.014e−02   4.607 4.08e−06 ***
langItalianTrue               −4.624e−02  5.796e−02  −0.798 0.424969
langSpanishTrue                9.468e−01  7.194e−02  13.161  < 2e−16 ***
langPortugeseTrue             −2.281e−01  8.913e−02  −2.560 0.010475 *
langGreekTrue                  4.187e−03  1.631e−01   0.026 0.979521
langCzechTrue                  3.085e−01  9.932e−02   3.106 0.001898 **
langSlovakTrue                −4.878e−01  1.075e−01  −4.537 5.71e−06 ***
langSlovenianTrue             −6.783e−01  1.473e−01  −4.606 4.11e−06 ***
langPolishTrue                 5.567e−01  1.711e−01   3.254 0.001139 **
langHungarianTrue              4.117e−01  9.996e−02   4.119 3.81e−05 ***
langRomanianTrue               3.695e−01  1.936e−01   1.909 0.056247 .
langBulgarianTrue              2.753e−01  1.728e−01   1.593 0.111202
langDanishTrue                 2.139e−01  1.515e−01   1.411 0.158183
langSwedishTrue                4.905e−01  1.382e−01   3.549 0.000387 ***
langFinnishTrue                3.895e−01  1.808e−01   2.155 0.031184 *
langLatvianTrue               −1.212e+00  8.624e−01  −1.405 0.159997
langEstonianTrue               2.894e−01  1.351e−01   2.142 0.032161 *
langLithuanianTrue            −7.331e−01  8.625e−01  −0.850 0.395345
lang_count                    −1.369e−01  2.170e−02  −6.309 2.80e−10 ***
auth1FATrue                    1.929e−03  3.637e−02   0.053 0.957694
auth2FATrue                   −2.074e−01  3.281e−02  −6.321 2.59e−10 ***
pop_score                      7.971e−07  2.954e−08  26.987  < 2e−16 ***
CountryBelgium                −2.713e−01  8.357e−02  −3.247 0.001168 **
CountryBulgaria               −5.882e−01  1.678e−01  −3.506 0.000455 ***
CountryCroatia                 5.577e−03  6.748e−02   0.083 0.934128
CountryCyprus                 −6.416e−01  1.823e−01  −3.519 0.000433 ***
CountryCzechia                 6.510e−01  1.062e−01   6.128 8.93e−10 ***
CountryDenmark                −1.151e−01  1.560e−01  −0.738 0.460465
CountryEstonia                        NA         NA      NA       NA
```

```
CountryFinland                   1.660e-02  1.734e-01   0.096 0.923726
CountryFrance                   -5.627e-01  5.654e-02  -9.953  < 2e-16 ***
CountryGermany                  -7.734e-01  3.397e-02 -22.768  < 2e-16 ***
CountryGreece                   -9.466e-01  1.699e-01  -5.572 2.51e-08 ***
CountryHungary                   5.615e-02  1.211e-01   0.464 0.642784
CountryIreland                  -5.111e-01  7.790e-02  -6.560 5.36e-11 ***
CountryItaly                    -1.569e-01  6.648e-02  -2.360 0.018282 *
CountryLatvia                    1.179e+00  8.673e-01   1.359 0.174004
CountryLithuania                 1.179e+00  8.673e-01   1.359 0.174004
CountryLuxembourg               -3.818e-01  6.788e-02  -5.625 1.86e-08 ***
CountryMalta                    -1.228e-01  3.951e-01  -0.311 0.756028
CountryNetherlands              -5.509e-01  8.356e-02  -6.593 4.32e-11 ***
CountryPoland                   -8.796e-01  1.706e-01  -5.157 2.51e-07 ***
CountryPortugal                  5.701e-01  1.047e-01   5.445 5.19e-08 ***
CountryRomania                   1.695e-01  1.914e-01   0.886 0.375804
CountrySlovakia                 -9.691e-04  1.011e-01  -0.010 0.992353
CountrySlovenia                  5.794e-01  2.081e-01   2.784 0.005365 **
CountrySpain                    -9.721e-01  7.502e-02 -12.957  < 2e-16 ***
CountrySweden                   -5.838e-01  1.398e-01  -4.176 2.96e-05 ***
CountryUnited Kingdom           -4.766e-01  5.153e-02  -9.249  < 2e-16 ***
threat_nameCitadel               4.146e+00  2.771e-01  14.963  < 2e-16 ***
threat_nameCoreBot               1.141e+00  7.360e-01   1.550 0.121116
threat_nameDridex-Loader         5.480e+00  2.767e-01  19.804  < 2e-16 ***
threat_nameDyre                  4.506e+00  2.783e-01  16.195  < 2e-16 ***
threat_nameGootkit               3.607e+00  2.768e-01  13.032  < 2e-16 ***
threat_nameGootkitLoader         2.107e+00  2.784e-01   7.569 3.76e-14 ***
threat_nameGozi-EQ               4.544e+00  2.789e-01  16.292  < 2e-16 ***
threat_nameGozi-ISFB             5.040e+00  2.770e-01  18.193  < 2e-16 ***
threat_nameIce9                  4.262e+00  2.862e-01  14.895  < 2e-16 ***
threat_nameKINS                  5.393e+00  2.775e-01  19.434  < 2e-16 ***
threat_nameKronos                3.029e+00  2.779e-01  10.901  < 2e-16 ***
threat_nameMatrix                1.815e+00  7.757e-01   2.339 0.019320 *
threat_nameNuclearBot            1.185e+00  2.780e-01   4.260 2.04e-05 ***
threat_nameNymaim                1.232e+00  1.057e+00   1.166 0.243749
threat_namePkybot               -7.069e-01  5.311e-01  -1.331 0.183199
threat_nameQadars                3.800e+00  2.771e-01  13.713  < 2e-16 ***
threat_nameQakbot                5.985e-01  1.174e+00   0.510 0.610269
threat_nameRamnit                2.128e+00  2.837e-01   7.499 6.42e-14 ***
threat_nameRamnit-BankerModule   2.090e+00  4.435e-01   4.712 2.45e-06 ***
threat_nameReactorBot            1.812e+00  5.443e-01   3.328 0.000874 ***
threat_nameRetefe-v2             4.241e+00  2.888e-01  14.687  < 2e-16 ***
threat_nameTheTrick              4.458e+00  2.777e-01  16.053  < 2e-16 ***
threat_nameTinba-v1              2.225e+00  2.805e-01   7.933 2.14e-15 ***
threat_nameTinba-v2              4.080e+00  2.770e-01  14.729  < 2e-16 ***
threat_nameZeuS                  4.770e+00  2.781e-01  17.155  < 2e-16 ***
threat_nameZeus-Action           3.865e+00  2.925e-01  13.212  < 2e-16 ***
threat_nameZeuS-OpenSSL          7.311e+00  2.768e-01  26.415  < 2e-16 ***
threat_nameZeuS-P2P              6.381e+00  2.806e-01  22.745  < 2e-16 ***
threat_nameZeus-Panda            4.229e+00  2.775e-01  15.239  < 2e-16 ***
```

```
year2015                      1.736e+00  2.455e-02  70.706  < 2e-16 ***
year2016                      7.935e-01  2.347e-02  33.805  < 2e-16 ***
year2017                      1.301e+00  2.501e-02  52.025  < 2e-16 ***
unique_attackurl_count        9.244e-03  2.960e-04  31.232  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.0333) family taken to be 1)

    Null deviance: 118979  on 33383  degrees of freedom
Residual deviance:  36173  on 33298  degrees of freedom
AIC: 313666

Number of Fisher Scoring iterations: 1


              Theta:  1.03331
          Std. Err.:  0.00737

 2 x log-likelihood:  -313492.18500
```

B.3. Negative Binomial Regression towards 7-day interval Attack Count

```
Call:
glm.nb(formula = week_attack_count ~ langEnglish + langGerman +
    langFrench + langDutch + langItalian + langSpanish + langPortugese +
    langGreek + langCzech + langSlovak + langSlovenian + langPolish +
    langHungarian + langRomanian + langBulgarian + langDanish +
    langSwedish + langFinnish + langLatvian + langEstonian +
    langLithuanian + lang_count + auth1FA + auth2FA + pop_score +
    Country + threat_name + year + unique_attackurl_count,
    data = data_notzero, init.theta = 1.622533869, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5714  -0.8813  -0.2708   0.2325  15.1667

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.560e+00  2.507e-01  -6.222 4.91e-10 ***
langEnglishTrue                1.235e-01  2.500e-02   4.941 7.75e-07 ***
langGermanTrue                -1.499e-01  3.210e-02  -4.668 3.04e-06 ***
langFrenchTrue                -2.843e-01  4.156e-02  -6.841 7.87e-12 ***
langDutchTrue                  3.644e-01  5.137e-02   7.094 1.31e-12 ***
langItalianTrue               -2.789e-01  4.898e-02  -5.694 1.24e-08 ***
langSpanishTrue                5.370e-01  6.100e-02   8.803  < 2e-16 ***
langPortugeseTrue             -6.766e-01  7.682e-02  -8.807  < 2e-16 ***
langGreekTrue                 -2.434e-01  1.424e-01  -1.710 0.087333 .
```

```
langCzechTrue                   -2.650e-02  8.510e-02   -0.311  0.755506
langSlovakTrue                  -5.051e-02  9.256e-02   -0.546  0.585255
langSlovenianTrue               -3.487e-01  1.251e-01   -2.787  0.005312 **
langPolishTrue                   1.747e-01  1.446e-01    1.209  0.226704
langHungarianTrue                6.233e-02  8.448e-02    0.738  0.460634
langRomanianTrue                 7.239e-01  1.646e-01    4.397  1.10e-05 ***
langBulgarianTrue                1.606e-01  1.495e-01    1.074  0.282707
langDanishTrue                   1.137e-01  1.288e-01    0.883  0.377438
langSwedishTrue                  3.180e-01  1.184e-01    2.687  0.007211 **
langFinnishTrue                  1.393e-01  1.542e-01    0.903  0.366327
langLatvianTrue                 -5.707e-01  9.031e-01   -0.632  0.527449
langEstonianTrue                -5.573e-01  1.189e-01   -4.685  2.80e-06 ***
langLithuanianTrue              -6.297e-01  9.032e-01   -0.697  0.485670
lang_count                       1.604e-02  1.857e-02    0.864  0.387858
auth1FATrue                     -1.504e-02  3.107e-02   -0.484  0.628313
auth2FATrue                     -1.733e-01  2.791e-02   -6.208  5.37e-10 ***
pop_score                        4.416e-07  2.516e-08   17.553   < 2e-16 ***
CountryBelgium                  -2.115e-01  7.130e-02   -2.966  0.003014 **
CountryBulgaria                 -4.521e-01  1.452e-01   -3.114  0.001844 **
CountryCroatia                  -1.857e-01  5.787e-02   -3.209  0.001334 **
CountryCyprus                   -5.356e-01  1.615e-01   -3.316  0.000914 ***
CountryCzechia                   4.342e-01  9.057e-02    4.795  1.63e-06 ***
CountryDenmark                  -1.959e-01  1.327e-01   -1.477  0.139720
CountryEstonia                          NA         NA       NA        NA
CountryFinland                  -1.784e-01  1.477e-01   -1.208  0.227237
CountryFrance                   -2.521e-01  4.827e-02   -5.223  1.76e-07 ***
CountryGermany                  -3.096e-01  2.907e-02  -10.649   < 2e-16 ***
CountryGreece                   -4.623e-01  1.476e-01   -3.132  0.001736 **
CountryHungary                   1.581e-01  1.021e-01    1.549  0.121441
CountryIreland                  -1.463e-01  6.565e-02   -2.228  0.025880 *
CountryItaly                    -3.646e-03  5.623e-02   -0.065  0.948306
CountryLatvia                    3.082e-01  9.068e-01    0.340  0.733946
CountryLithuania                 3.082e-01  9.068e-01    0.340  0.733946
CountryLuxembourg                2.310e-02  5.786e-02    0.399  0.689718
CountryMalta                    -3.742e-01  3.755e-01   -0.997  0.318996
CountryNetherlands              -4.581e-01  7.125e-02   -6.430  1.28e-10 ***
CountryPoland                   -6.368e-01  1.440e-01   -4.421  9.81e-06 ***
CountryPortugal                  2.221e-01  9.020e-02    2.462  0.013817 *
CountryRomania                  -3.233e-01  1.630e-01   -1.984  0.047273 *
CountrySlovakia                 -3.020e-01  8.734e-02   -3.457  0.000546 ***
CountrySlovenia                  2.293e-02  1.793e-01    0.128  0.898224
CountrySpain                    -6.121e-01  6.411e-02   -9.547   < 2e-16 ***
CountrySweden                   -4.172e-01  1.197e-01   -3.487  0.000489 ***
CountryUnited Kingdom           -3.232e-01  4.376e-02   -7.387  1.50e-13 ***
threat_nameCitadel               2.288e+00  2.449e-01    9.344   < 2e-16 ***
threat_nameCoreBot               5.620e-01  6.751e-01    0.832  0.405143
threat_nameDridex-Loader         3.799e+00  2.445e-01   15.538   < 2e-16 ***
threat_nameDyre                  3.135e+00  2.457e-01   12.761   < 2e-16 ***
threat_nameGootkit               2.415e+00  2.446e-01    9.874   < 2e-16 ***
```

```
threat_nameGootkitLoader          9.520e−01  2.467e−01   3.859 0.000114 ***
threat_nameGozi−EQ                2.619e+00  2.463e−01  10.634  < 2e−16 ***
threat_nameGozi−ISFB              3.594e+00  2.447e−01  14.686  < 2e−16 ***
threat_nameIce9                   2.040e+00  2.532e−01   8.057 7.81e−16 ***
threat_nameKINS                   2.965e+00  2.452e−01  12.092  < 2e−16 ***
threat_nameKronos                 2.097e+00  2.456e−01   8.538  < 2e−16 ***
threat_nameMatrix                 2.161e−01  7.327e−01   0.295 0.768016
threat_nameNuclearBot             6.094e−01  2.461e−01   2.477 0.013267 *
threat_nameNymaim                 1.068e+00  8.717e−01   1.225 0.220482
threat_namePkybot                −1.574e+00  5.113e−01  −3.079 0.002078 **
threat_nameQadars                 1.968e+00  2.450e−01   8.031 9.68e−16 ***
threat_nameQakbot                −5.242e−01  1.295e+00  −0.405 0.685654
threat_nameRamnit                 1.098e+00  2.513e−01   4.370 1.24e−05 ***
threat_nameRamnit−BankerModule    1.262e+00  3.859e−01   3.270 0.001075 **
threat_nameReactorBot             7.411e−01  4.955e−01   1.496 0.134737
threat_nameRetefe−v2              1.286e+00  2.562e−01   5.020 5.18e−07 ***
threat_nameTheTrick               3.228e+00  2.453e−01  13.162  < 2e−16 ***
threat_nameTinba−v1               1.067e+00  2.493e−01   4.282 1.86e−05 ***
threat_nameTinba−v2               2.849e+00  2.448e−01  11.640  < 2e−16 ***
threat_nameZeuS                   2.791e+00  2.457e−01  11.361  < 2e−16 ***
threat_nameZeus−Action            6.100e−01  2.634e−01   2.316 0.020552 *
threat_nameZeuS−OpenSSL           3.630e+00  2.446e−01  14.844  < 2e−16 ***
threat_nameZeuS−P2P               3.677e+00  2.476e−01  14.854  < 2e−16 ***
threat_nameZeus−Panda             2.389e+00  2.452e−01   9.739  < 2e−16 ***
year2015                          1.100e+00  2.076e−02  52.976  < 2e−16 ***
year2016                          2.607e−01  2.028e−02  12.858  < 2e−16 ***
year2017                          6.794e−01  2.163e−02  31.409  < 2e−16 ***
unique_attackurl_count            1.087e−02  2.449e−04  44.386  < 2e−16 ***
−−−
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for Negative Binomial(1.6225) family taken to be 1)

    Null deviance: 84802  on 33383  degrees of freedom
Residual deviance: 32131  on 33298  degrees of freedom
AIC: 216737

Number of Fisher Scoring iterations: 1


              Theta:  1.6225
          Std. Err.:  0.0135

 2 x log−likelihood:  −216563.4640
```

B.4. Negative Binomial Regression towards Unique Attack ID Count

```
Call:
glm.nb(formula = id_attack_count ~ langEnglish + langGerman +
    langFrench + langDutch + langItalian + langSpanish + langPortugese +
    langGreek + langCzech + langSlovak + langSlovenian + langPolish +
    langHungarian + langRomanian + langBulgarian + langDanish +
    langSwedish + langFinnish + langLatvian + langEstonian +
    langLithuanian + lang_count + auth1FA + auth2FA + pop_score +
    Country + threat_name + year + unique_attackurl_count,
    data = data_notzero, init.theta = 1.767001301, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
−3.9040  −0.6102  −0.1625   0.2683  11.4126

Coefficients: (1 not defined because of singularities)
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  −4.946e−01  2.454e−01  −2.016 0.043834 *
langEnglishTrue               1.931e−01  2.544e−02   7.590 3.19e−14 ***
langGermanTrue                4.822e−03  3.277e−02   0.147 0.883021
langFrenchTrue               −9.358e−02  4.218e−02  −2.219 0.026518 *
langDutchTrue                 3.957e−01  5.181e−02   7.638 2.20e−14 ***
langItalianTrue              −2.563e−01  4.995e−02  −5.131 2.88e−07 ***
langSpanishTrue               7.938e−01  6.095e−02  13.025  < 2e−16 ***
langPortugeseTrue            −8.685e−01  7.881e−02 −11.020  < 2e−16 ***
langGreekTrue                 1.816e−01  1.455e−01   1.248 0.212025
langCzechTrue                 4.732e−02  8.721e−02   0.543 0.587400
langSlovakTrue                4.928e−02  9.546e−02   0.516 0.605659
langSlovenianTrue            −4.320e−01  1.351e−01  −3.197 0.001386 **
langPolishTrue                1.703e−01  1.438e−01   1.184 0.236409
langHungarianTrue             1.579e−03  8.600e−02   0.018 0.985346
langRomanianTrue              7.145e−01  1.680e−01   4.252 2.12e−05 ***
langBulgarianTrue             1.708e−01  1.513e−01   1.129 0.258784
langDanishTrue                4.537e−02  1.285e−01   0.353 0.723981
langSwedishTrue               3.792e−01  1.177e−01   3.223 0.001269 **
langFinnishTrue               3.299e−01  1.536e−01   2.147 0.031764 *
langLatvianTrue               8.135e−02  7.989e−01   0.102 0.918888
langEstonianTrue             −2.490e−01  1.227e−01  −2.028 0.042512 *
langLithuanianTrue           −2.012e−02  7.990e−01  −0.025 0.979910
lang_count                   −2.782e−03  1.900e−02  −0.146 0.883602
auth1FATrue                  −3.747e−02  3.150e−02  −1.190 0.234206
auth2FATrue                  −1.064e−01  2.816e−02  −3.779 0.000158 ***
pop_score                     3.636e−07  2.566e−08  14.169  < 2e−16 ***
CountryBelgium               −2.739e−01  7.330e−02  −3.737 0.000186 ***
CountryBulgaria              −3.293e−01  1.464e−01  −2.249 0.024491 *
CountryCroatia               −1.481e−02  5.969e−02  −0.248 0.804077
CountryCyprus                −3.486e−01  1.670e−01  −2.088 0.036784 *
CountryCzechia                3.153e−01  9.274e−02   3.399 0.000676 ***
CountryDenmark                6.379e−02  1.322e−01   0.482 0.629511
CountryEstonia                       NA         NA      NA       NA
CountryFinland                1.466e−01  1.461e−01   1.004 0.315450
CountryFrance                −6.957e−02  4.922e−02  −1.413 0.157540
```

```
CountryGermany                 -4.475e-02  2.984e-02  -1.500 0.133605
CountryGreece                  -1.195e-01  1.511e-01  -0.791 0.429107
CountryHungary                  3.159e-01  1.034e-01   3.054 0.002259 **
CountryIreland                  1.089e-01  6.613e-02   1.646 0.099706 .
CountryItaly                   -1.093e-01  5.748e-02  -1.902 0.057178 .
CountryLatvia                  -1.758e-01  8.034e-01  -0.219 0.826822
CountryLithuania               -1.758e-01  8.034e-01  -0.219 0.826822
CountryLuxembourg               1.371e-01  5.867e-02   2.337 0.019443 *
CountryMalta                    1.252e-01  3.664e-01   0.342 0.732492
CountryNetherlands             -2.335e-01  7.165e-02  -3.258 0.001121 **
CountryPoland                  -3.769e-02  1.435e-01  -0.263 0.792794
CountryPortugal                 5.311e-01  9.313e-02   5.703 1.18e-08 ***
CountryRomania                 -2.226e-01  1.664e-01  -1.338 0.180871
CountrySlovakia                -1.563e-01  9.032e-02  -1.730 0.083606 .
CountrySlovenia                 3.100e-01  1.913e-01   1.621 0.105046
CountrySpain                   -5.444e-01  6.517e-02  -8.354  < 2e-16 ***
CountrySweden                  -2.828e-01  1.188e-01  -2.380 0.017305 *
CountryUnited Kingdom          -1.469e-01  4.445e-02  -3.304 0.000953 ***
threat_nameCitadel              1.146e+00  2.393e-01   4.790 1.67e-06 ***
threat_nameCoreBot              8.650e-02  6.570e-01   0.132 0.895246
threat_nameDridex-Loader        2.515e+00  2.388e-01  10.532  < 2e-16 ***
threat_nameDyre                 3.814e+00  2.400e-01  15.893  < 2e-16 ***
threat_nameGootkit              1.730e+00  2.389e-01   7.242 4.41e-13 ***
threat_nameGootkitLoader        6.176e-01  2.409e-01   2.564 0.010353 *
threat_nameGozi-EQ              1.503e+00  2.407e-01   6.243 4.30e-10 ***
threat_nameGozi-ISFB            3.905e+00  2.389e-01  16.344  < 2e-16 ***
threat_nameIce9                 2.122e-01  2.496e-01   0.850 0.395207
threat_nameKINS                 2.346e+00  2.395e-01   9.797  < 2e-16 ***
threat_nameKronos               1.775e+00  2.398e-01   7.399 1.37e-13 ***
threat_nameMatrix              -5.130e-01  7.694e-01  -0.667 0.504866
threat_nameNuclearBot           8.166e-01  2.400e-01   3.403 0.000667 ***
threat_nameNymaim              -1.061e+00  1.060e+00  -1.001 0.316865
threat_namePkybot              -1.645e+00  5.002e-01  -3.288 0.001009 **
threat_nameQadars               1.566e+00  2.393e-01   6.545 5.96e-11 ***
threat_nameQakbot              -6.011e-01  1.274e+00  -0.472 0.637139
threat_nameRamnit               8.668e-01  2.458e-01   3.527 0.000420 ***
threat_nameRamnit-BankerModule  1.181e+00  3.688e-01   3.202 0.001366 **
threat_nameReactorBot           1.361e+00  4.437e-01   3.067 0.002160 **
threat_nameRetefe-v2            2.498e+00  2.477e-01  10.084  < 2e-16 ***
threat_nameTheTrick             3.715e+00  2.394e-01  15.520  < 2e-16 ***
threat_nameTinba-v1             1.424e+00  2.419e-01   5.887 3.93e-09 ***
threat_nameTinba-v2             8.064e-01  2.393e-01   3.370 0.000751 ***
threat_nameZeuS                 2.329e-01  2.407e-01   0.968 0.333225
threat_nameZeus-Action          2.327e-01  2.588e-01   0.899 0.368588
threat_nameZeuS-OpenSSL         2.303e+00  2.389e-01   9.642  < 2e-16 ***
threat_nameZeuS-P2P             8.582e-01  2.428e-01   3.534 0.000409 ***
threat_nameZeus-Panda           2.355e+00  2.394e-01   9.839  < 2e-16 ***
year2015                       -1.282e-01  2.220e-02  -5.775 7.70e-09 ***
year2016                       -5.313e-01  2.184e-02 -24.332  < 2e-16 ***
```

```
year2017                       −3.456e−01  2.290e−02 −15.095  < 2e−16 ***
unique_attackurl_count          1.214e−02  2.433e−04  49.872  < 2e−16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for Negative Binomial(1.767) family taken to be 1)

    Null deviance: 117087  on 33383  degrees of freedom
Residual deviance:  29822  on 33298  degrees of freedom
AIC: 185590

Number of Fisher Scoring iterations: 1


              Theta:  1.7670
          Std. Err.:  0.0158
Warning while fitting theta: alternation limit reached

 2 x log−likelihood:  −185415.5960
```

B.5. Summary of Regression Models

Table B.1: Summary of logistic model and negative binomial model towards different metrics

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
langEnglishTrue	0.629^{*}	0.379^{***}	0.124^{***}	0.193^{***}
	(0.314)	(0.029)	(0.025)	(0.025)
langGermanTrue	0.515	0.265^{***}	-0.150^{***}	0.005
	(0.385)	(0.038)	(0.032)	(0.033)
langFrenchTrue	0.355	0.010	-0.284^{***}	-0.094^{*}
	(0.455)	(0.049)	(0.042)	(0.042)
langDutchTrue	0.604	0.277^{***}	0.364^{***}	0.396^{***}
	(0.561)	(0.060)	(0.051)	(0.052)

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
langItalianTrue	0.847	−0.046	-0.279^{***}	-0.256^{***}
	(0.639)	(0.058)	(0.049)	(0.050)
langSpanishTrue	−0.061	0.947^{***}	0.537^{***}	0.794^{***}
	(0.625)	(0.072)	(0.061)	(0.061)
langPortugeseTrue	0.885	-0.228^{*}	-0.677^{***}	-0.868^{***}
	(1.022)	(0.089)	(0.077)	(0.079)
langGreekTrue	1.160	0.004	−0.243	0.182
	(1.211)	(0.163)	(0.142)	(0.146)
langCzechTrue	0.097	0.308^{**}	−0.026	0.047
	(1.106)	(0.099)	(0.085)	(0.087)
langSlovakTrue	2.075	-0.488^{***}	−0.051	0.049
	(1.074)	(0.108)	(0.093)	(0.095)
langSlovenianTrue	−1.987	-0.678^{***}	-0.349^{**}	-0.432^{**}
	(1.817)	(0.147)	(0.125)	(0.135)
langPolishTrue	−1.573	0.557^{**}	0.175	0.170
	(1.623)	(0.171)	(0.145)	(0.144)
langHungarianTrue	1.987	0.412^{***}	0.062	0.002
	(1.561)	(0.100)	(0.084)	(0.086)
langRomanianTrue	−17.352	0.370	0.724^{***}	0.714^{***}

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
	(391.660)	(0.194)	(0.165)	(0.168)
langBulgarianTrue	11.961	0.275	0.161	0.171
	(1,484.644)	(0.173)	(0.149)	(0.151)
langDanishTrue	−0.543	0.214	0.114	0.045
	(1.197)	(0.152)	(0.129)	(0.128)
langSwedishTrue	2.146^{**}	0.490^{***}	0.318^{**}	0.379^{**}
	(0.793)	(0.138)	(0.118)	(0.118)
langFinnishTrue	−1.894	0.389^{*}	0.139	0.330^{*}
	(1.335)	(0.181)	(0.154)	(0.154)
langLatvianTrue	0.248	−1.212	−0.571	0.081
	(1.568)	(0.862)	(0.903)	(0.799)
langEstonianTrue	-3.439^{***}	0.289^{*}	-0.557^{***}	-0.249^{*}
	(0.916)	(0.135)	(0.119)	(0.123)
langLithuanianTrue	15.970	−0.733	−0.630	−0.020
	(2,399.545)	(0.862)	(0.903)	(0.799)
lang_count	−0.035	-0.137^{***}	0.016	−0.003
	(0.258)	(0.022)	(0.019)	(0.019)
auth1FATrue	0.458	0.002	−0.015	−0.037
	(0.286)	(0.036)	(0.031)	(0.032)

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
auth2FATrue	2.148^{***}	-0.207^{***}	-0.173^{***}	-0.106^{***}
	(0.266)	(0.033)	(0.028)	(0.028)
pop_score	0.00000^{***}	0.00000^{***}	0.00000^{***}	0.00000^{***}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
CountryBelgium	-1.844^{**}	-0.271^{**}	-0.211^{**}	-0.274^{***}
	(0.702)	(0.084)	(0.071)	(0.073)
CountryBulgaria	14.129	-0.588^{***}	-0.452^{**}	-0.329^{*}
	(1,369.951)	(0.168)	(0.145)	(0.146)
CountryCroatia	0.797	0.006	-0.186^{**}	−0.015
	(0.806)	(0.067)	(0.058)	(0.060)
CountryCyprus	-1.879^{*}	-0.642^{***}	-0.536^{***}	-0.349^{*}
	(0.869)	(0.182)	(0.162)	(0.167)
CountryCzechia	−0.351	0.651^{***}	0.434^{***}	0.315^{***}
	(1.204)	(0.106)	(0.091)	(0.093)
CountryDenmark	-4.040^{**}	−0.115	−0.196	0.064
	(1.265)	(0.156)	(0.133)	(0.132)
CountryEstonia				
CountryFinland	1.552	0.017	−0.178	0.147
	(1.358)	(0.173)	(0.148)	(0.146)

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
CountryFrance	−0.678	−0.563***	−0.252***	−0.070
	(0.550)	(0.057)	(0.048)	(0.049)
CountryGermany	−0.095	−0.773***	−0.310***	−0.045
	(0.299)	(0.034)	(0.029)	(0.030)
CountryGreece	−2.406	−0.947***	−0.462**	−0.119
	(1.277)	(0.170)	(0.148)	(0.151)
CountryHungary	−5.971***	0.056	0.158	0.316**
	(1.576)	(0.121)	(0.102)	(0.103)
CountryIreland	−3.671***	−0.511***	−0.146*	0.109
	(0.512)	(0.078)	(0.066)	(0.066)
CountryItaly	−0.898	−0.157*	−0.004	−0.109
	(0.703)	(0.066)	(0.056)	(0.057)
CountryLatvia	−1.435	1.179	0.308	−0.176
	(1.623)	(0.867)	(0.907)	(0.803)
CountryLithuania	−18.188	1.179	0.308	−0.176
	(2,399.545)	(0.867)	(0.907)	(0.803)
CountryLuxembourg	−2.082***	−0.382***	0.023	0.137*
	(0.529)	(0.068)	(0.058)	(0.059)
CountryMalta	−3.517***	−0.123	−0.374	0.125

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
	(1.032)	(0.395)	(0.376)	(0.366)
CountryNetherlands	−1.343	−0.551***	−0.458***	−0.233**
	(0.694)	(0.084)	(0.071)	(0.072)
CountryPoland	−0.145	−0.880***	−0.637***	−0.038
	(1.632)	(0.171)	(0.144)	(0.143)
CountryPortugal	−6.810***	0.570***	0.222*	0.531***
	(1.073)	(0.105)	(0.090)	(0.093)
CountryRomania	16.885	0.169	−0.323*	−0.223
	(391.660)	(0.191)	(0.163)	(0.166)
CountrySlovakia	−1.136	−0.001	−0.302***	−0.156
	(1.035)	(0.101)	(0.087)	(0.090)
CountrySlovenia	−0.733	0.579**	0.023	0.310
	(2.003)	(0.208)	(0.179)	(0.191)
CountrySpain	0.351	−0.972***	−0.612***	−0.544***
	(0.702)	(0.075)	(0.064)	(0.065)
CountrySweden	−3.840***	−0.584***	−0.417***	−0.283*
	(0.867)	(0.140)	(0.120)	(0.119)
CountryUnited Kingdom	−1.060*	−0.477***	−0.323***	−0.147***
	(0.444)	(0.052)	(0.044)	(0.044)

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
threat_nameCitadel		4.146***	2.288***	1.146***
		(0.277)	(0.245)	(0.239)
threat_nameCoreBot		1.141	0.562	0.087
		(0.736)	(0.675)	(0.657)
threat_nameDridex-Loader		5.480***	3.799***	2.515***
		(0.277)	(0.245)	(0.239)
threat_nameDyre		4.506***	3.135***	3.814***
		(0.278)	(0.246)	(0.240)
threat_nameGootkit		3.607***	2.415***	1.730***
		(0.277)	(0.245)	(0.239)
threat_nameGootkitLoader		2.107***	0.952***	0.618*
		(0.278)	(0.247)	(0.241)
threat_nameGozi-EQ		4.544***	2.619***	1.503***
		(0.279)	(0.246)	(0.241)
threat_nameGozi-ISFB		5.040***	3.594***	3.905***
		(0.277)	(0.245)	(0.239)
threat_nameIce9		4.262***	2.040***	0.212
		(0.286)	(0.253)	(0.250)

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
threat_nameKINS		5.393^{***}	2.965^{***}	2.346^{***}
		(0.278)	(0.245)	(0.239)
threat_nameKronos		3.029^{***}	2.097^{***}	1.775^{***}
		(0.278)	(0.246)	(0.240)
threat_nameMatrix		1.815^{*}	0.216	−0.513
		(0.776)	(0.733)	(0.769)
threat_nameNuclearBot		1.185^{***}	0.609^{*}	0.817^{***}
		(0.278)	(0.246)	(0.240)
threat_nameNymaim		1.232	1.068	−1.061
		(1.057)	(0.872)	(1.060)
threat_namePkybot		−0.707	-1.574^{**}	-1.645^{**}
		(0.531)	(0.511)	(0.500)
threat_nameQadars		3.800^{***}	1.968^{***}	1.566^{***}
		(0.277)	(0.245)	(0.239)
threat_nameQakbot		0.599	−0.524	−0.601
		(1.174)	(1.295)	(1.274)
threat_nameRamnit		2.128^{***}	1.098^{***}	0.867^{***}
		(0.284)	(0.251)	(0.246)

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
threat_nameRamnit-BankerModule		2.090***	1.262**	1.181**
		(0.443)	(0.386)	(0.369)
threat_nameReactorBot		1.812***	0.741	1.361**
		(0.544)	(0.495)	(0.444)
threat_nameRetefe-v2		4.241***	1.286***	2.498***
		(0.289)	(0.256)	(0.248)
threat_nameTheTrick		4.458***	3.228***	3.715***
		(0.278)	(0.245)	(0.239)
threat_nameTinba-v1		2.225***	1.067***	1.424***
		(0.280)	(0.249)	(0.242)
threat_nameTinba-v2		4.080***	2.849***	0.806***
		(0.277)	(0.245)	(0.239)
threat_nameZeuS		4.770***	2.791***	0.233
		(0.278)	(0.246)	(0.241)
threat_nameZeus-Action		3.865***	0.610*	0.233
		(0.293)	(0.263)	(0.259)
threat_nameZeuS-OpenSSL		7.311***	3.630***	2.303***
		(0.277)	(0.245)	(0.239)

	Dependent variable:			
	is_targeted	raw_attack_count	week_attack_count	id_attack_count
	logistic	*negative binomial*	*negative binomial*	*negative binomial*
	(1)	(2)	(3)	(4)
threat_nameZeuS-P2P		6.381***	3.677***	0.858***
		(0.281)	(0.248)	(0.243)
threat_nameZeus-Panda		4.229***	2.389***	2.355***
		(0.278)	(0.245)	(0.239)
year2015		1.736***	1.100***	−0.128***
		(0.025)	(0.021)	(0.022)
year2016		0.794***	0.261***	−0.531***
		(0.023)	(0.020)	(0.022)
year2017		1.301***	0.679***	−0.346***
		(0.025)	(0.022)	(0.023)
unique_attackurl_count		0.009***	0.011***	0.012***
		(0.0003)	(0.0002)	(0.0002)
Constant	−0.856	−2.704***	−1.560***	−0.495*
	(0.486)	(0.284)	(0.251)	(0.245)
Log Likelihood	−1,845.873	−156,747.100	−108,282.700	−92,708.800
θ		1.033*** (0.007)	1.623*** (0.014)	1.767*** (0.016)
Akaike Inf. Crit.	3,795.746	313,666.200	216,737.500	185,589.600
Note:				*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets

C
Expert Interview Protocol

As the interview is in semi-structured manner, the real question given to the interviewees might differ, depending on the flow of the interview, but will still be in line with the general protocol.

C.1. Standard Protocol

- Introduce yourself, the background and the purpose of this interview session
- Ask for permission to record the audio of the conversation process
- Ask the questions in the orderly manner

C.2. Background & Purpose of the Interview

The purpose of this thesis is to gather information and then analyze whether certain characteristics of online banking could explain its target selection, i.e. whether they are likely to be targeted or not targeted. More specifically, the language offered by the online banking, the authentication factor it applied, and its domain popularity are assessed.

The interview will consist of 2 phases. The first phase will discuss more about the general perspective about factors that could affect the target selection. The discussion will be more focused later on the factors that are assessed in this thesis. The second phase will try to get an interpretation from experts regarding the statistical model and result generated in this thesis. The purpose is to collect expert opinion regarding the result that arises from the model as well as possible explanation and/or interpretation about the relationship that are shown in the model.

C.3. Questions

[Begin the interview: Introduction]

For the protocol, could you tell us your name, your institution/company name and your role in the institution/company?

Phase 1: Getting general perspective

1. In your opinion, what are factors of banks or their online banking services that could affect the target selection, i.e. could explain why certain online banking is more/less/not targeted by the banking malware?

- Could you explain how the factors that you mentioned affect the target selection?

2. Let's discuss the factors besides the financial ones

   - (If any non-financial factor has not been mentioned) Is there any non-financial factor that you think is also significant to affect the target selection?
     - Could you explain how this/these factor(s) affect the target selection?
     - How important is/are this/these factor(s) in explaining the target selection, especially if compared to the financial factors?
   - (If any non-financial factor was mentioned) You mentioned some of non-financial factors, how important is/are this/these factor(s) in explaining the target selection, especially if compared to the financial factors?

3. This thesis focuses on assessing certain characteristics of online banking, such as languages offered, authentication factor applied, and the domain popularity.

   - In general, how do you think the language could affect the target selection? Could you elaborate more on that?
   - Focused on 2 big languages in Europe: English and German, how do you think these languages could affect the target selection?
   - Regarding the authentication factor, how do you think it could affect the target selection, i.e. would applying 2-factor authentication would make any difference compared to applying 1-factor authentication?
   - Regarding the domain popularity, how do you think it could affect the target selection, i.e. would more popular domain lead to the online banking being more/less/not targeted?

Phase 2: Getting interpretation about the model result

[Present a document showing the result of the regression model and initial interpretation is given]

1. What are your opinion about the result?

2. What do you think the reason behind the relationships presented in the result?

3. What do you think about the model and result in general? Do you have any concern?

4. What do you think about the importance of this (kind of) model in getting more understanding about the target selection of banking malware in banking sector?

5. What do you think could make the model better in the future?

[End of Interview]

D

Transcriptions of the Expert Interviews

D.1. The Interview with Paul Samwel

The interview was conducted on 18 June 2018 in Rabobank office, Utrecht.

Q: Could you tell me your name and your institution and your role in the institution?

A: My name is Paul Samwel. I work for Rabobank, Head of Security Architecture and Innovation and Cyber Threat Intelligence.

Q: In your opinion, what are factors of banks or their online banking service that could affect the target select?

A: The most important factor, I guess, is something to gain for the criminals, that is, do we have a positive business case. The business case was created back in the 90s when we start with online banking. Once there is something to gain, they start looking for the easier target. That can be a result of controls and also language issue. In the first phase of online banking fraud we definitely see that English-speaking countries were more attacked than other languages. In a short point of time, we also saw a huge spike of malware-related attacks in the Netherlands. We feel it is related to the fact that we have a small banking community in the Netherlands. We only have three banks which account for almost 90 percent of market. It means if you have a botnet with 1000 infected PCs and a malware that can successfully target one of these banks then you can use 90 percent of the botnet. In contrary, in German there are many smaller banks so it is harder as you need many different types of malware.

Q: You mentioned that positive business case is the case and it is related to financial factors...

A: Yes. The best example I can give is the Dutch tax institute. They had certain ways of authentication and security controls that worked fine for years and no criminal activities at all. Before, people only needed to pay taxes. Then, they also used the same administration to pay people toeslagen [translation: subsidy]. From that moment on, they started to see crime. It is because no one wants to pay someone's taxes but they want to receive someone's subsidies. Nothing changed in terms of controls but suddenly they saw frauds.

Q: Do you see other non-financial factors besides the one you mentioned like language?

A: The kind of control you have, the effectiveness of the attack. When we introduced security monitoring, it became harder for them to get the money and they started to shift to other countries. Another example, when we started to be very strict to money mules, by exchanging account numbers of money mules between other Dutch banks and starting a campaign in Dutch schools to educate students about money mules, they definitely had an impact on the fraud levels.

Q: How important are the non-financial factors compared to the financial factors for this case, in your opinion?

A: I think criminals make a calculation so the financial factors of the business case of the attack is the most important. They calculate how much money they have to invest and what do they want to get out from it and they should make sure that the outcome is positive. They also compare that with doing the same job in other countries or areas. We also have a feeling that malware attacks do not only shift from a country to another country, but also they might use the same infrastructure, the same botnet to do another kind of fraud, like distributing ransomware. In the end, financial aspect is the most important one and other factors like language or authentication factor are something that you can calculate into finance. Back in the past, between 2000-2005, we did not see any attack on Dutch banks; we see lots of attacks on UK banks. It can partially be a language issue, but also at that time, US and UK banks still used user ID and password as an authentication mechanism. The phishing technique to get user ID and password is much easier than the malware you need to change transactions that use multi-factor authentication.

Q: In general, how do you think the language can affect the target selection?

A: If you look at phishing mails, I do think they have a capability of understanding our website. You might use Google translate and that works fine to get the feeling on what the website is about. But that is not enough anymore to mislead people. To social engineer people, you need similar to proper Dutch, not google-translated Dutch. Therefore, they started to hire people with those languages. I also have a personal feeling they trick people who are in the Netherlands but do not speak good Dutch. I think in the past, about 10 years ago, I saw many phishing mails with very poor Dutch and I think it is kind of deliberate. The first step in attack is a cheap step, just sending mails. If you send mails in poor Dutch, a small percentage, 2 or 3 percent, of people will fall in the social engineering scheme, and those people are very easy to be convinced in the more expensive second step because they are the not smart one. We do not see this anymore because we explained so much time to all people about this.

Q: How do you think English and German affect the target selection? You have explained about English before.

A: We feel that malware is done with criminal groups outside EU, therefore for most of them, both Dutch and German are foreign languages. But, of course more people speaks German and it is easier to find someone that can help you with German. That is not a big problem there.

Q: How do authentication factor affect the target selection, in your opinion?

A: I think a good form of authentication is a good way to make the attack less successful, but nowadays you need something more like digital signature or display the account number and the amount in a way they cannot change. It helps a lot of clients to stop the fraudulent transactions. If you are able to stop 95 percent of potential fraudulent transactions, then it helps changing the business case. The problems are those kinds of controls are rather expensive and the clients might not find it very user friendly because they needed a separate device.

Q: How do domain popularity affect the target selection, in your opinion?

A: For malware type of attacks, they need to invest in malware that is able to interact with the dialog of the bank. Therefore, it helps if it is a large bank, so you have a lot of victims to get your investment back. For other social engineering like by phone, it does not matter whether you are small or large bank. With the spreading of the malware, it also helps if you have a large relative market share. Usually they buy infected machines and, as I said in the beginning, if you have, for example, 30 percent of market share, you are an interesting victim. Back in the time we had a huge spike, you saw that only 3 banks in the Netherlands suffered a majority of the fraud. Other banks are not even attacked, and they have less strict control. We believed that at that time there were criminal organization from Eastern Europe that bought so many infected machines and they might only invest in writing malware for ABN Amro, ING and Rabobank and not bother with other smaller banks. For instance, Van Landschot, a private bank with very rich customers, they only had a market share of about 2 percent and it was simply not interesting enough for the criminals. But nowadays, we see different types of attacks like human social engineering not involving malware and with those frauds, Van Landschot is suffering the same kind of frauds.

Q: [Present the analysis result]

A: It is a bit dangerous to explain all the differences you see because you forgot the full scope of all attacks and try to explain them with only two hypotheses: the language and the authentication factor. There might be more factors, like, I got a feeling that relative market share is more important than language. When frauds started in Europe, we started to explain the French model. In France they used to use weak authentication, but at that time you can only transfer money to certain number of accounts. In order to use another account, you need to do a manual process to add the account. It was a different type of control, but in France, it was very effective that time as we did not see any fraud in France. It was adopted by Dutch banks by having higher limit for transfer to account it has transferred money to and lower limit to those you do not know before. They are some other factors that might also be relevant and perhaps there are five other factors that could explain the differences. But now you are trying to explain the differences you see with only two factors. They might not be the real reason. The relative market share and the language might have a correlation, but maybe the language is not the issue, but the relative market share is the issue. Maybe you can look at the same data with more factors and you could see different results.

Q: Regardless of other factors that become the limitation of this model, what do you think about the result or the insights?

A: I think you cannot draw a conclusion from the model because you consider a limited number of factors. In 2013, when we had a huge spike in internet frauds and at that time we have two-factor authentication and only offer Dutch. With your result you might conclude that the combination of Dutch and two-factor authentication makes them very interesting. But, another factor that makes us very interesting that time is that we are the only country in Europe that had more than 90 percent of people doing online banking. In France, at that time, the majority of people still do banking using paper. If you look at Africa, for instance, they use totally different techniques. Another thing, if you compare the average bank account in Poland and in the Netherlands, for instance, there are more money on it so that is more interesting for criminals. There are more factors to explain the differences. In a more proper research, I would suggest to add 10 or 15 more factors and see which one has an influence and which not.

D.2. The Interview with Maarten Jak

The interview was conducted on 28 June 2018 in ABN Amro office, Amsterdam.

Q: Could you tell me your name and your institution and your role in the institution?

A: My name is Maarten Jak. I have been for 6 years in ABN Amro. I studied in TU Delft with a minor in Security, Safety and Justice. That makes me interested to go into cybercrime. My role in ABN Amro is also in the intelligence team in cybersecurity field.

Q: In your opinion, what are factors of banks or their online banking service that could affect the target select?

A: In the beginning of the malware concept, there was a lot of automatic malware. Basically, it scanned the banks and saw if the English-looking page is present. Sometimes it worked, sometimes it did not. Now they are more using sophisticated approach in which the malware is not so much automated because a lot of banks are now improving their detection. With a lot of bank improving their fraud detection, the criminals became more sophisticated in the target selection as it involved more manpower. They really made a specific malware for a specific bank and when it is triggered, it was not automatically ingested to the session, but the real criminals have to make a transaction via a backdoor or whatever to make it look like that it is from the client so that the fraud detection is not triggered. How they did the target selection? I think they still look on the English page since English is still a quite big language. But we also see they attack Dutch pages because they just hire a local in the country which can translate a page for them. So, they are local groups; we have Dutch malware, Belgium malware, German malware and so on. Whether two-factor authentication is also a factor, basically almost all two-factor authentication can be circumvented nowadays, especially by the malware coming in a couple of years because they just ask it, for example, in an Android “security update” infecting the phone or they ask it directly in the page itself, for example, by showing that there is an error and you need to click to external page. I do not know the most specific reason for the target selection, probably whether it works or not. We saw a lot of tests on a couple of banks and if it worked we see quite a lot of infection. I think, based on these tests, they made a decision to attack these banks and infect all of their customers in this country. Also, another one is how many banks are there in a specific location. If there are a lot of banks in a big nation, it is quite hard to get the right infection at the right people, because there are so many banks that one person can have and you do not know whether your payloads will work or not.

If you look at the graph of online malware and the loss it created in the Netherlands, we can see the number dropped from 16 millions to 11 millions in a couple of months so we are quite good at training our detection system to tackle the malware. You also see that criminals are going to the more conventional methods again instead of online banking malware because it does not work again anymore, or they go to other countries where the security is less.

Q: For the factors you mentioned before, which one is the most important?

A: The number of customers of a bank is quite important because, as I said, if there are a lot of banks in a nation, it will be quite hard to infect the right person. Here in the Netherlands, we have only 4 great banks, so it is quite easy to target someone: there is 25 percent chance that he or she is part of those banks. I think that is a big factor. For the language, as I said, they just hire people who can translate the malware for them so there are many local variances of malware in local languages. I do not think English is a great influencer, but of course every big bank has English page so that is the first selection of finding the banks or how big they are, so it makes sense that it is a factor. And also security: if the test on a bank fails, they might move on to other banks.

Q: You described the influence of English, how about another big language in Europe like German?

A: The problem with Germany is that they have more than 600 small banks. It is quite hard to write 600 different types of malware to get the right customers of the right bank. But it is up and coming. The malware is less automated now, but more people powered so they do not have to write a complete automated transaction flow but it just need to takeover the session and put a loading script in front of it, so it is becoming easier to target a big nation with a lot of banks. To answer whether the language is influential, in a common sense it is, but I think they look more towards how big the bank is, how many customers and the number of banks in that area. Also, Germany is not that far in online banking as we are. Coming back to the previous question, this might also be a factor: how people are used to the online banking. In the Netherlands, we are quite used to do all the transactions online and that is why there is a high success rate.

Q: How do authentication factor affect the target selection, in your opinion?

A: By applying two-factor authentication, you have more security about if it is the right person that did the transaction. The two-factor authentication should be in a hand of the actual person who opened the account. Of course, there are many ways to circumvent it. But it should give more security about the person. However, it is indeed not the only defense in a whole detection system; it is just one layer on top of other layers of defense strategy. But one-factor authentication is not enough anymore because nowadays you can take leaked user ID and password anywhere. It does not give any sense of security. However, for your information, there is a company claiming that they have very good detection system so that you should not have to use any authentication form at all.

Q: How do domain popularity affect the target selection, in your opinion?

A: If it is higher in popularity, it can become more popular target because it should have more customers, more money and higher density in the country itself. That makes it more rewarding to write a specific malware for it.

Q: [Present the analysis result]

A: I saw the languages with negative coefficient refer to countries that have low density rate of online banking environment. So, writing a specific malware for those countries do not give much reward because not every bank there has online banking and not many people are used to it. In the Netherlands we have 89 percent online banking density so writing malware here is much more rewarding as there is more chance people will click.

I also see that the data is only the malware set from Fox IT so there has already been a shift in the malware that you can analyze. There might be other malware starting to infect the banks for other reasons, but you do not have the dataset for that to be included in your variables. You then make assumptions based on the limited dataset that you have. It is good to remind yourself of that if you make conclusion, you make the conclusion based on the dataset that you have, and it is not the same as all malware that targets the banks out there. It could be bias.

Q: Regardless of other factors that become the limitation of this model, what do you think about the result or the insights?

A: It is good to look at the malware samples and try to identify why do they target some banks, because if we know the reason, we could incorporate it somewhere in the defense mechanism and help other banks. The result itself is not that surprising, it is quite logical and still in line with the other discussion in this field. But, it is nice to confirm your initial thoughts.

Q: What could make the model be improved?

A: More data.

D.3. The Interview with Huub Roem

The interview was conducted on 25 July 2018 in ING Bank office, Amsterdam.

Q: Could you tell me your name and your institution and your role in the institution?

A: My name is Huub Roem, I am a forensic IT expert at the cyber defense center within ING. We are serving domestic banks in the Netherlands and we are investigating e-frauds and doing incident response in case of internal threats.

Q: In your opinion, what are factors of banks or their online banking service that could affect the target selection?

A: First of all, language, in terms of software development, but also language of the user interface.

Q: Is there any other factor?

A: The market share, I think, is also important, because if you are a small bank with few customers, then you are not so international, you are not specific to the target. So, a small bank in the Netherlands, with only a few thousand customers, is not interesting for global actors.

Q: So, you implied that banks with a large market share will be targeted more?

A: Yes, that is correct.

Q: Anything else?

A: The complexity of the authentication methods, and things like that.

Q: You mentioned about language before. Could you explain or elaborate more how this factor could influence the target selection?

A: Yes. I think Dutch is a foreign actor, difficult to learn and difficult to handle within social engineering component of the malware, so it is very important for the target selection. To be more precise, for about 1.5 years, myING was provided only in Dutch and we saw actors moving away from myING.

Q: Is it because they do not understand Dutch so they found another target that provides another language?

A: Yes, or to succeed is too complex due to the language, because you need to create a web inject in perfect Dutch without any misspell.

Q: You also mentioned about the authentication method. Could you explain or elaborate more?

A: Within the Netherlands, we are the only bank using SMS authentication. For the development of the actors and the malware, maybe it is more difficult to design a malware and to act with such tools. Other Dutch banks use tokens and when it is more general to work with tokens, I think they will avoid ING for creating another version of malware.

Q: So, did you say that SMS authentication is harder for the attacker to bypass than token authentication?

A: I don't think so. But, because it is different from other banks, they have to develop some other mechanism to act.

Q: I see. So, you have mentioned some financial factors and also some non-financial factors...

A: Maybe there is another factor. When you have the [system] development which is going very quickly, that is, you make a lot of changes to your service, myING in our case, then it is also related. If there is a lot of changes, the actor has to change the malware as well. When you have a relevant change for avoiding the malware every 2 or 3 weeks, I think the actor would move away from you.

Q: How do you see a presence of a language, like English or German, in terms of the target selection of malware?

A: It is significant because if you look at the current malware, TrickBot is one of the largest malware this time, for the last 2 years they only attack the bank with English language, for example, they only target the US, UK and Australia. We also see some development within, for example TrickBot, meaning we are now in its config, after we introduced English in myING. For German, I think it is just another important language in Europe. In terms of potential victims, the Germans are interesting for actors because the Germans have a lot of money.

Q: How do different authentication factor affect the target selection, in your opinion?

A: Criminal always look at the easiest way and it's much easier to grab only a password than grab a password and a token.

Q: How do domain popularity affect the target selection, in your opinion?

A: I think it affects. Especially for criminals outside our region. Maybe the criminals use marketing tools to select the target.

Q: [Present the analysis result]

A: Do you look at the timestamp of the malware and the attack period? Because we see a lot of ING within the configuration of the malware but in fact they don't really attack our customers. Sometimes they are using addons to the config to do some analytics. When they only see, for example, 10 infections out of 10 thousand, they simply skip us in the next configuration or they don't even develop a web inject to target our customers. So, ING could be within the config, but most of the times there are no web inject available for ING. They are using it just for analytics. It could be relevant for selecting the final target.

Q: Thank you for the input. Besides, in general, what do you think about the result?

A: It surprises me, because now I get that the characteristics are not relevant for selecting the target.

Q: Do you have any concern regarding the model?

A: I think, in real life, the decision to select the target is more complex than those. It is now based on only 4 and 5 items. In real life, it's more complex.

Q: What do you think about the importance of this kind of model in understanding the target selection?

A: I think, of course. For the organization, you can include this in the risk model to make a relevance in your risk score.

Q: What do you think could make the model better in the future?

A: Adding more data to it.

Another thing, at this moment, we see a slight decrease of the use of malware. Law enforcements have caught a lot of important criminals, banks in Europe are more mature, better in detection, making the success rate of automated and hybrid malware lower and lower. Therefore, they are moving away from a specific malware and acting more with the social engineering component using social media and all kind of channels to target the customers, outside of our banking channel.