WHO IS NEXT?

Identifying characteristics of European banks that are key in influencing the target selection of banking malware.

M.E. Hoppenreijs

University of Technology Delft

Who is next?

Identifying characteristics of European Banks that are key in influencing the target selection of banking malware

by

M.E. Hoppenreijs

Master thesis submitted to Delft University of Technology in partial fulfilment of the requirements for the degree of

Master of Science
in Engineering and Policy Analysis

To be defended in public on March 21, 2019

Student number:	4243986	
Project duration:	September 3, 2018 – March 21, 2019	
Thesis committee:	Prof. dr. ir. M.J.G. van Eeten,	TU Delft, chair
	Dr. ir. C.H. Gañán,	TU Delft, 1st Supervisor
	Dr. M.E. Warnier,	TU Delft, 2nd Supervisor
	D. Perk,	Fox-IT, External Supervisor

Preface

This Master Thesis is a result of six months of research as final examination of a master degree in ”Engineering and Policy Analysis” at the Delft University of Technology. This study aims to provide insight in which banks’ characteristics increase the likelihood of being targeted by banking malware. It is intended for anyone who wants to seek insight into the target selection process of banking malware, such as cybersecurity experts in academics, in the financial and private sector.

This research process has been a learning experience. I have gained considerable knowledge of banking malware, enlarged my Python skills and even worked with advanced data analytic methods. A few highly knowledgeable people have supported me throughout the process which I highly appreciated. I would like to express my gratitude for all their efforts to make this research to a success.

First of all, I would like to thank my supervisory team of the TU Delft: Prof. dr. M.J.G. (Michel) van Eeten, Dr. ir. C.(Carlos) Hernandez Gañán and Dr. M. (Martijn) Warnier. Michel van Eeten, thank you for giving me the opportunity to research this topic. Carlos Hernandez Gañán, you have been involved from the start of this project. I would like to thank you for the many e-mail conversations, for helping me to acquire the data, and checking in on me once in a while. Martijn Warnier, I fully understand why you are such a popular committee member for many of our students. You are able to quickly grasp the typical student’s struggles and you certainly don’t need many words to understand them. Thank you for being such an engaged supervisor when I voiced my concerns throughout the project.

In addition, I would like to thank Fox-IT and the Threat Intelligence department of Fox-IT for having me as a graduate intern. In particular, I would like to thank Diederik Perk for the Monday afternoon sessions, it was always helpful to bat ideas about my research around you. I also benefited from debating my results with the analyst of the department sharing their knowledge of years of reverse engineering of banking malware with me.

Samuel Natalius, many thanks for sharing the clear structured data and resources with me. Your patience and help in answering my questions was superb. My gratitude also goes to the experts in this research for sharing their knowledge, experience and opinion in the interview sessions.

On a more personal note, I would like to thank my parents and brother for supporting me throughout the university years and continuously reading my drafts, Peter for facilitating a no-stress zone, my friends and volleyball team for their advice and support.

I hope you will enjoy reading this thesis and find it interesting to take a deep dive into the criminal mind of banking malware deployers.

M.E. Hoppenreijs
Delft, March 2019

Executive Summary

The European financial sector is a frequent victim of banking malware leading to massive losses. Banking malware is aimed to commit financial fraud or theft by gaining access to a customer's bank account. It appears that not all customers' banks are equally attractive targets among cybercriminals who deploy banking malware. From all targets, 20 % of the banks attract over 80 % of the targets (Natalius, 2018; Tajalizadehkhoob, Asghari, Gañán, & Eeten, 2014). There is a limited understanding of why certain banks are more attractive to cybercriminals than others. Understanding the underlying reasons and trends behind the banking malware attacks can support decision-making about technical and policy-based mitigation and may even reduce the impact of the effects of banking malware attacks. Therefore, this research aims to answer the following question; **Which characteristics of European banks are key in influencing the target selection process of banking malware?**

The objective of this research is to establish a comprehensive model explaining the target selection of banking malware. Earlier research made an effort to develop such a model but was not able to include financially related characteristics, and those are, according to the experts in the field essential, to explain why a bank is targeted (Natalius, 2018). For this reason, the goal is to include bank-size in the explanatory model and explore its relationship with target selection. Multiple bank-size measurements are selected and after extensive literature research, the most suitable bank-size measures are chosen: revenues, equity, total assets, market capital, net income, risk-weighted assets, number of (online) customers, number of employees, branches, loans and deposits of customers. These measurements are combined into one **bank-size** variable by performing a Principal Component Analysis.

This research analyses if the following characteristics of European banks influence target selection: *bank-size, domain-popularity, brand-value, language-use on websites, being part of a banking group, and two-factor authentication.* Data on those variables is collected from previous research and open-source data. This dataset is combined with a dataset containing information of targeted banks over the year 2016-2017. Ultimately, the merged dataset contains information for 1293 European banks, 887 targeted banks and 406 non-targeted banks for the year 2016 and 2017. The data is analysed by performing two types of regression analysis. First, logistic regression is conducted to get insight into the key banks' characteristics influencing whether a bank has been targeted. Second, a negative binomial regression is performed to identify the critical features of banks explaining the frequency of being targeted. Experts validated the regression results and provided additional insight into the target selection process.

Endogeneity and highly correlated banks' characteristics in the regression model might affect the estimated coefficients of the regression models. It is, therefore, difficult to draw valid conclusions based on the output of the regression model. Nevertheless, the results of the regression model along with the insights gained by the expert interviews, provide some very profound knowledge on the banks' characteristics in target selection.

Experts acknowledged the importance of **bank-size** in target selection, but they also envision a trend in targeting smaller banks because they have inferior security measures compared to the larger banks. However, proven by the descriptive and regression analysis, larger banks have a higher probability of being targeted and are a more frequent target of banking malware. The above statement is not false but needs some nuance and explanation. Based on the knowledge gained in this research, cybercriminals focusing on certain countries and expand their focus by targeting more various banks. This has as a consequence that also smaller banks are being targeted. It can be concluded, that larger banks remain an

appealing target for cybercriminals but also smaller banks are targeted since they have less fraud prevention to obtain ”ease money”.

Visibility-related factors, i.e. **domain popularity, brand value and being part of a banking group** are positive predictors for banks to be targeted and also to be more frequently targeted. These visibility factors are stronger predictors for target selection compared to bank-size. This can be reasoned by the bounded rationality of Cybercriminals Simon (1972). Criminals have limited information, cognitive limitation, and a limited amount of time to make a decision. Research on bank robberies shows that criminals select banks which they hoped to retrieve the largest financial rewards, but those were not always accurate predictions (Morrison, 1996). It is likely that similar patterns are visible in the digital world. In this way, banks that are visible have a higher chance to be targeted because they have higher expected rewards according to the cybercriminals.

The presence of the **two-factor authentication** has a positive influence on being targeted, but it negatively influences the frequency of being targeted. This controversial relation can be explained as follows. Apparently, two-factor authentication is not a barrier for cybercriminals since they can circumvent the security measure by using man-in-the-browser attack, such as spyware/spyrootkits (Adham, Azodi, Desmedt, & Karaolis, 2013) or man-in-the-middle attacks (Mitnick, 2017). The reduction of targets as a result of having two-factor authentication can be reasoned by the effectiveness of certain types of two-factor authentication or to complementary security measures that correspond with having two-factor authentication.

Various experts emphasise that the use of the primarily spoken **languages** on the banks' website eases the attack. At the start of the banking malware, banks having interfaces that offer largely spoken languages, i.e. English and German were more targeted since it is less challenging to copy these interfaces. At this moment, interfaces in all languages have been seen and language is no longer a barrier. This argumentation is in line with the results of the analysis, various - not largely spoken languages influenced target selection. The cause of the importance of these languages is mostly related to size, popularity or the level op two-factor adaptation in a country.

In line with the outcomes of this research, some recommendations are proposed. It is recommended to put the bank-size metric and the domain-popularity into practice when performing risk assessments. The bank-size metric and domain popularity (a strong predictor for being targeted) can be used to assess potential losses and attack probabilities for risk analysis or even used to specify (cyber) insurance premiums. For smaller banks, it is recommended to invest in rule-based detection and customer awareness when allocating scarce resources since experts recognised a huge decrease when implementing rule-based engines. Larger banks should be critical towards allowing digital customer registration or allowing multiple bank accounts for a single customer. These services are necessary for larger banks to increase customers experiences, but they also make the bank prone to banking malware attacks.

This thesis aspires to empower research analysts to enhance and improve the current regression model. Improving the model can be done by performing multiple principal component analysis to combine banks characteristics to reduce independence in the model, for example, the visibility variables. Also, add control variables related to the social context of a country to eliminate false-correlation. As an extension of this research, potential steps are to investigate the effectiveness of the various types of two-factor authentication. Furthermore, some additional data and features can be added to the regression model, such as, add more yearly data, add features related to the ease of transferring illegal money and investigate the influence of banks operating at (inter)national level.

Contents

List of Figures

List of Tables

Abbreviations

APT	Advanced Persistent Threats
AUC	Area Under the Curve
C&C	Command and Control
DGA	Domain Generation Algorithm
DNS	Domain Name System
EBA	European Bank Authority
ECB	European Central Bank
ECDF	Empirical Cumulative Distribution Function
EPA	Engineering and Policy Analysis
IQR	Interquartile Range
MFI	Monetary Financial Institutions
MitB	Man-in-the-Browser
MitM	Man-in-the-Middle
P2P	Peer-to-Peer
PCA	Principal Component Analysis
RFE	Recursive Feature elimination
RAT	Routine Activity Theory
ROC	Receiver Operating Characteristic
RWA	Risk-Weighted Assets
SMB	Service Message Block
SMOTE	Synthetic Minority Over Sampling Technique
TTP	Tactics Techniques and Procedures
VIF	Variance Inflation Factor
VNC	Virtual Networking Computer

1

Introduction

1.1. Banking malware in the financial sector

Since the introduction of online banking, financial institutions have been a target for cybercrime. Particularly, consumer fraud is one of the leading factors for financial losses since the discovery of banking malware Zeus in 2006. Ever since, new threats emerge utilising advances techniques. Sophisticated attacks are deployed by skilled organised groups, such as Carbanak and Dridex, stealing billions of dollars from bank networks (Carter, 2017).

Banking malware, a category of financial malware, is an information-stealing malicious software aimed at committing financial fraud or theft. Banking malware has remained active in the threat landscape for many years. In the past years, banking malware attacks have evolved in frequency, volume and sophistication (Chebyshev, Sinitsyn, Parinov, Liskin, & Kupreev, 2018). Significant incidents will continue to take place as it stays profitable for cybercriminals to conduct these attacks (Allisy-Roberts et al., 2018; Wueest, 2017).

Today's world is marked by its complex geopolitical environment, and new technologies continue to reshape financial services. Banks have the challenging task to secure their online transactions in this interlinked and globalised world. Although banks have improved their security measures, cybercriminals adapted the tactics and techniques of their attacks and exploited the accessibility of the online services (Allisy-Roberts et al., 2018). Dealing with strategic behaviour of multiple (individual or organised) actors in a dynamic environment denotes the complexity for banks to protect themselves against banking malware attacks.

It appears that not all customers' banks are equally attractive for cybercriminals to target. Proven by quantitative research of Tajalizadehkhoob et al. (2014) and Natalius (2018), 80% of all attack attempts target 20% of all banks. Nonetheless, there is a limited understanding of why certain financial institutions are more likely to be selected as a target. Gathering intelligence about the underlying reasons, trends, methods and capabilities of the banking malware attacks could help banks to gain a better understanding regarding their adversaries' activities and make informed decisions about the technical and policy-based mitigation strategies to reduce targeted attacks (Miles, Lakhotia, LeDoux, Newsom, & Notani, 2014; Shackleford, 2015).

This research aims to identify which characteristics of European banks determine the attractiveness for cybercriminals. It aims to seek insight into which bank features define the attractiveness of that particular bank. The analysed banks' characteristics form the explanatory model of target selection. Banks characteristics of previous scholars' work with a similar dataset and geographical focus will be added to develop a more complete explanatory model and, consequently, improve the quality of the model and research. The established model supports shaping effective defence strategies.

1.2. Literature review

Target selection is a popular research topic for understanding criminal decision-making in physical crime (robbery) and terrorism. It is also an upcoming topic to get a better understanding of the digital crime. This section describes the routine activity theory as theoretical framework to understand online fraud, in particular, malware attacks. In addition, this section explores the fields of current target selection research. This literature review identifies the knowledge gap for the next section 1.3, that will further shape the direction of this research.

1.2.1. Routine Activity Theory

Most researchers try to understand and identify the cause of online victimisation using criminology theories. Routine Activity Theory (RAT) proves its success in accounting for the traditional physical crime and tries to apply this to online crimes. Pratt, Holtfreter, and Reisig (2010) found that RAT contributes to a theoretically informed direction for online fraud prevention. The research of Jaishankar (2011) examines how malware infection (crimes that do not exist in physical time and space) can be addressed by the RAT. According to those researchers, a burglary that can be explained by RAT has similarities with a malware infection. The findings of this research provide partial support for the application of RAT to data loss caused by malicious software. Both authors advocate a continued examination of the RAT framework for cybercrime through empirical analysis.

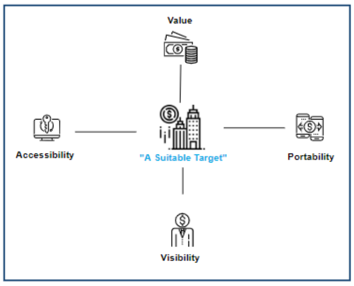


Figure 1.1: Routine Activity Theory (RAT)

Yar (2005) argued that RAT is rather limited in explaining cybercrime because of its socio-international characteristics, i.e. internet connectivity, anonymity, spatial-temporal barriers. He introduced four fundamental properties that define a “suitable target”, namely: *value, portability, visibility, and accessibility* (Figure 1.1). Value stands for how much financial gain can be achieved. Portability is the ease at which money can be transferred to the adversaries account. Visibility relates to how visible a target is towards adversary. Accessibility is how easy the target can be reached.

Various target selection studies, such as Tajalizadehkhoob et al. (2014); Van Moorsel (2016) and Natalius (2018) draw upon the RAT theory. Natalius (2018); Tajalizadehkhoob et al. (2014) used the three aspects of the RAT’s suitable target element: value, visibility and accessibility to identify characteristics that influence the target selection of banking malware attacks.

Leukfeldt and Yar (2016) argued that some RAT elements are not able to adequately explain victimisation in cybercrime, such as the financial characteristics that play a role in making the decision to deploy banking malware. The analysis shows that economic aspects, such as personal income, household income, financial assets and financial possessions of victims do not increase the risk for identity theft or malware theft. This phenomenon can be explained by the fact that malware aims to target as many individuals as possible, and therefore this research states that the explanation of victimisation shifts from ’value’ and ’visibility’ towards ’online accessibility’.

1.2.2. Target selection in bank robbery

The key determinants of target selection in bank robberies are extensively investigated and described below. Bank robberies and digital attacks differ obviously in nature, but the strong link between the physical threat and the digital threat provides opportunities to understand the cybercriminals mind.

Wright and Decker (2002) state that target selection is an important element in understanding the robbery process. Multiple researchers studied what the key determinants of bank robbers were in choosing a specific bank to rob. This research was done by interviewing criminals and analysing police reports. Most of the robbers mainly focused on a banks interior (hiding places, getaways and security measures) and did not feel at risk nor were they afraid of being caught (Weisel, 2007; Wright & Decker, 2002). The robbers were even unconcerned about alarms and cameras. They also chose their targets based on the gains they hoped to achieve (Morrison, 1996; Wright & Decker, 2002). Banks with the highest financial expectations were more likely to be robbed, but often those expectations did not meet reality. The research of Morrison (1996) showed that the (un)availability of weapons was rarely an indicator in a robbers assessment since it is relatively easy to obtain (fake) weapons.

Concluding, the decision-making process of bank robbery heavily relies on the financial rewards the criminal hopes to achieve. Clear risks are accepted or simply ignored, and resources are abundant at hand. Comparing this with banking malware targeting, it is possible that similar reasons, financial gain and easy access, underlie target selection. However, for digital crime a high level of technical knowledge and time is required to conduct a successful attack; resources really matter. Thus, in contrast to physical crime, resources and capabilities of the adversaries may certainly play an essential role in the decision-making process.

1.2.3. Target selection in cyber crime

Target selection in the cyber domain is a modern research topic. The definition of target selection deviates among researches. Van Moorsel (2016) defined target selection specific for banking malware: the attack choices by the financial malware schemes regarding which financial institution to attack, at which point of time and for how long. This research applies a more general definition of target selection by The MITRE Corporation (2017).

> *" Target selection is the iterative process of an adversary to determine a target, starting at a strategic level and narrow it down operationally and tactically until a specific target is chosen"*

Target selection is part of the reconnaissance step (first step) of the cyber kill chain. The research of target selection for banking malware started with Tajalizadehkhoob et al. (2014). She researched target selection by analysing Zeus configurations files. In her first research, she discovered that banks with the English language on their banking web pages are attacked more within the EU region. Also, countries with a higher rate of GDP and broadband penetration were attacked more globally. She proceeded her research by studying the value of US banks based on their deposits. According to her, size is a threshold factor but does not predict the intensity of the attacks. Cheung (2017) researched target selection from DDoS amplification attacks. This research used factors on a country level, namely the ICT development index and GDP. He proved that an increase in ICT development index and GDP resulted in a decreased number of attacks. Van Moorsel (2016) investigated the relation between the number of clients and the attack intensity in ten European countries. This research showed that only in 2009, there is a significant relation observed where more clients increase the attack intensity. Beazley Breach (2016) proved that hacking and malware attacks increasingly targeted smaller and more vulnerable financial institutions, particularly those with annual revenues of under $35 million.

Europol (2017) has an explanation of why certain banks are more attractive to target for banking malware. They argue that criminals prefer to deal with a limited number of bank accounts where they can get as much as possible. Knowledge on when the detection system alarms the transaction as fraudulent is essential in order to bypass the security control systems. Criminals, therefore, prefer to deal with similar bank accounts to get as much as possible payments. Florêncio and Herley (2010) state that easiness to recruit money mules is the explanation of the repetition to target specific banks.

The four elements of a 'suitable target'- RAT theory, will be used to provide a structured overview of all researched characteristics in the current scientific research (see Table 1.1). It distinguishes two types of scientific research: the quantitative analysis (first row) and the qualitative analysis (second row). The qualitative study expresses the knowledge and expertise gathered by interviews or (cyber) reports, whilst the quantitative studies used data-analytics to verify their expectation.

Table 1.1: Overview of analysed banks' characteristics in current target selection literature

Value	Visibility	Accessibility	Portability
Quantitative analyse			
Bank-size			
Number of (online) customers (Tajalizadehkhoob, 2013; Van Moorsel, 2016)	Domain name visibility* (Tajalizadehkhoob, 2013)	Bank authentication method (Natalius, 2018)	
Revenues (Beazley Breach, 2016)	Website domain popularity** (Tajalizadehkhoob, 2013)	Broadband penetration rate (Tajalizadehkhoob, 2013)	
At country Level			
Financial Status (GDP) (Natalius, 2018)	Language of the online banking *** (Cheung, 2017)		
ICT Development Index (Cheung, 2017; Natalius, 2018)			
Rate of banking penetration (Van Moorsel, 2016)			
Qualitative Analyse			
Degree of cooperation between financial institutions and law enforcement within a country (Van Moorsel, 2016)	Brand popularity (Kaspersky, 2016)	Users online awareness (Tajalizadehkhoob, 2013; Van Moorsel, 2016)	Ease in securely performing criminal actions (transferring money) (Wueest, 2017)
Bank Size: number of (online) customers, total assets, total payments, revenues, net profit, relative market share (Natalius, 2018)	Banks attack record (Tajalizadehkhoob, 2013; Van Moorsel, 2016)	Degree of collaboration between (big) financial institution on cybersecurity level within a country (Van Moorsel, 2016)	Money transfer policies of the country (Tajalizadehkhoob, 2013; Van Moorsel, 2016)
Country: financial status, number of internet users and rate of banking/shopping penetration (Natalius, 2018)	Ownership of the bank (public/private) (Van Moorsel, 2016)	Security control (firewalls and the quality of detection)(Van Moorsel, 2016)	Ease of recruiting money mules (Florêncio & Herley, 2010; Natalius, 2018)
Adaption rate of online banking (Natalius, 2018)		Maturity of online banking systems (Natalius, 2018)	

Note: * not significant; **weak significant; ***significant

In a nutshell, adversaries choose a target based on multiple factors, e.g. expected financial gain, popularity, the maturity of security measures, and the ease of monetising. Another potential factor for criminals is the availability of resources, such as the IT-infrastructures (back-end systems or the development of malicious bank websites). It is most likely that a combination of defenders and attacker characteristics explains target selection in financial malware fraud. This research will focus on the features of the defender; thus it will identify which characteristics of specific banks increases the attractiveness for attackers. Information on which banks are targeted will be extracted from the configuration files of the banking malware, what this means will be extensively explained in Chapter 2.

1.3. Knowledge gap and Research Objective

The following knowledge gaps are identified based on the literature review in the previous section.

Explanatory model on target selection of banking malware

While target selection studies in Table 1.1, offer essential insight in target selection for banking malware, they do not provide a comprehensive overview of target selection. For example, the research of Van Moorsel (2016) showed that the number of clients influenced the frequency of the attack, while experts argue that the country caused this increase in attack frequency. In addition, the size of a bank can be expressed by more factors than the number of clients. A comprehensive explanatory model, including many variables and determinants, is essential to reduce false correlations and get a complete understanding of target selection. Since the aforementioned researches relied on a limited sampling set (Zeus malware), focus on a single type of attack (DDoS) or differ in geographical focus, it is impossible to combine the researched variables into one explanatory model explaining target selection in banking malware. Only the research of Natalius (2018) analysed various banking malware families. To manifest a complete understanding of target selection and to include all malware families, this research paper will continue and extend the work of previous scholars (Natalius, 2018). Natalius (2018) created an explanatory model of three characteristics, namely: two-factor authentication, language-use and domain popularity.

Bank-size in the cyber domain

According to experts from the field, the financial and market features are missing in the research of Natalius (2018). This research aims to add the missing financial and market indicators and develop a valid explanatory model for target selection. Although, the relation between bank-size and systemic risk is a traditional research topic among economists, the relation between bank-size and cyber risk, such as banking malware, is very limited researched (Laeven, L., Ratnovski, L., & Tong, 2014). Cheung (2017) investigated if organisation size correlates with the number of DDoS attacks. This analysis limited itself by only utilising Fortune500 data. Tajalizadehkhoob (2013) investigated the value of the US banks based on their deposits and Van Moorsel (2016) investigated the bank-size regarding the number of bank clients for ten European nations. Yet, measuring bank-size, next to deposit and the number of clients, can be done by a wide selection of factors. Schildbach and Schneider (2017) show that it is important to not rely solely on a single measure to evaluate bank-size because it can influence the outcome. For example, US banks are larger compared to European banks, if you take market capital as the only indicator. On the other hand, European banks are larger than US banks if you take total assets or revenues as indicator (Schildbach & Schneider, 2017).

This research analyses the effect of bank-size, using a wide variety of factors related to the probability of being targeted by banking malware. It identifies, analyses and adds the bank-size to an explanatory model.

Trends in target selection

After defining bank-size and whether it influences the target frequency, the trend for targeting bigger or smaller financial institutions will be investigated. According to the literature, adversaries tend to target smaller financial institutions, since the bigger banks traditionally work more closely together and have improved their security. Smaller financial institutions do not always have the resources to combat the attacks (Beazley Breach, 2016; Van Moorsel, 2016). However, if they have them, the effects can be considerable. An example is the strong decrease of card-not-present fraud in 2011 in The Netherlands. This decline could be attributed to several preventive measures taken by the banks like geo-blocking, physical protecting of ATM's against skimming, transaction monitoring and close cooperation with the police (Dutch Banking Association, 2017). Geo-blocking is a security feature that restricts the use of cards in some parts of the world. The trend of targeting smaller financial institutions has never been verified by scientific research but will still be covered in this research.

Contribution to RAT

Leukfeldt and Yar (2016) argue that some RAT elements are not able to adequately explain victimisation in cybercrime such as the financial characteristics that play a role in decision-making to deploy banking malware. The analysis shows that financial characteristics, such as personal income, household income, financial assets and monetary possessions of victims do not affect the risk of becoming a victim of identity theft or malware. Malware targets as many individuals as possible and malware is not able to discriminate or even identify financial gain. Therefore, Leukfeldt and Yar (2016) see that victimisation shifts from 'value' and 'visibility' towards 'online accessibility'. This research focused on the characteristics of the individual (customers) and did not evaluate the characteristics of the bank that might play a part in target selection. This research paper contributes to the RAT if the theory is able to explain victimisation by banking malware toward European banks.

Research Objectives

The research objectives are defined based on the above-mentioned knowledge gaps.

1. Explore the relationship between bank-size and target selection.
   - Identify if smaller institutions are a more frequent victim of banking malware.
   - Contribute to RAT by exploring if RAT is applicable to explain victimisation.
2. Developing an explanatory model for target selection in order to generate a comprehensive overview of which factors are key for an adversary to select certain banks when deploying malware. The following factors will be analysed: bank-size, language-use on the banks' website, domain-popularity, brand-value, being part of a banking group and the two-factor authentication.

1.4. Research questions and approach

The research objective in the previous section aims to seek a comprehensive understanding of how cybercriminals select their targets. To achieve this, the following research question will be answered in this study:

Which characteristics of European banks influence target selection of banking malware?

A design approach will assure structure towards answering the research question. The research proposes to follow four phases based on the explanatory mixed method design by Creswell and Clark (2011). This design follows a quantitative approach followed by a qualitative approach.

Phase 1, the conceptualisation phase examines all relevant scientific theories and describes the threat landscape of banking malware such as the threat actors or common attack vectors. Phase 2 is the quantitative data collection phase. This phase collects external data necessary for the regression analyse, i.e. the bank-size of financial institutions. In Phase 3, the quantitative analysis performs a regression analysis and develops the explanatory model of target selection. Finally, in Phase 4, the qualitative data collection and analysis, conduct expert interviews and validates the results of the regression model.

In line with the four research phases of the design approach, the following sub-questions can be defined. The sub-research questions as presented in this paragraph will logically answer the primary research question.

1. *Conceptualisation.* How does the current banking malware threat landscape look like?

2. *Quantitative data collection and metrics identification.* How is bank-size delineated in current literature and how can bank-size be measured?

3. *Quantitative analysis.* Do bank-size, language-use on bank' website, domain popularity, brand value, being part of banking group and two-factor authentication influence whether a bank is targeted or the intensity of the targets?

4. *Qualitative data collection and analysis.* Which analysed banks' characteristics are essential in target selection according to experts?

1.5. Research methodology

This paragraph describes the selected methods and their limitations: A) Desk research, B) Statistical data analysis, and C) Expert validation.

(A) **Desk research**
For the first and second phase of the research, desk research will be conducted. It foresees a literature review to generate a full understanding of how financial malware performs in the threat landscape. The required data for this phase could be gathered from libraries, journals and databases (e.g. Web of Science, Scopus or Science Direct). Phase 2 collects bank-size based indicators from external resources. The limitation of a literature review in the field of target selection is that the concepts are largely unpublished. Furthermore, the dynamic digital environment and the sophistication of the malware attacks are constantly changing. For this reason, a complete, reliable, and 100 per cent accurate literature review may be difficult to establish. This limitation also applies to the external data collection since a reliable and complete (mostly classified) dataset is difficult to obtain or to use for research purposes.

(B) **Statistical data analysis**
The third phase, quantitative data collection, refers to statistical data analysis. This research paper aims to provide intelligence on the adversaries target selection process. To generate intelligence, data has to be analysed and correlated by using regression analysis. The analysis uses a database provided by Fox-IT that extracts information from the configuration files of the malware. The outcomes of the regression will be the basis for an explanatory model. Explanatory research is an instrument to identify the ratio behind a wide range of processes. A limitation of this method is that correlations sometimes turn out to be accidental, which indicates a false-correlation. As a consequence, incorrect conclusions could be drawn. It is therefore important to include relevant characteristics of target selection and to identify the relationship.

(C) **Expert validation**
In Phase 4, the qualitative data collection and analysis, expert interviews will be conducted to obtain critical opinions on the target selection problem. There is a lack of literature or intelligence about target selection, and therefore the quantitative analysis of the previous phase cannot be validated by literature, and thus expert interviews are necessary to review the model. Potential limitations of expert validation are the lack of available experts, data-bias, loss of focus and scope of the research.

1.5.1. Research flow and structure

Figure 1.2 provides a visual representation of the research flow. This diagram is a work-flow guide to the research question. It shows the phases of the research (light-grey; defined in Section 1.4), the methodology (dark-blue; defined in Section 1.5), and the upcoming chapters (orange).

This research is structured as follows:

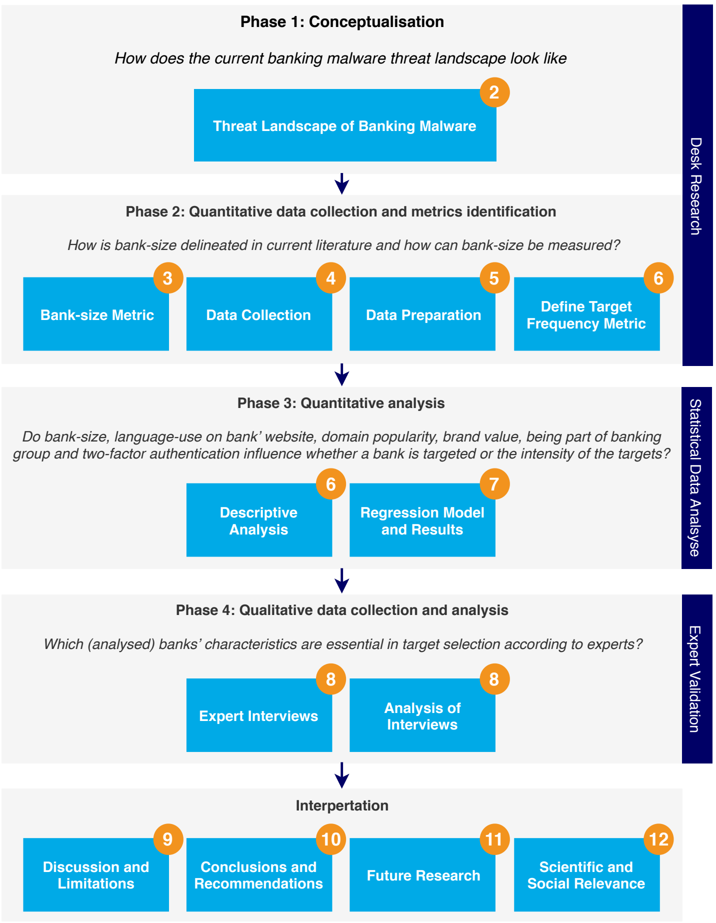

Phase 1: Conceptualisation
How does the current banking malware threat landscape look like
2
Threat Landscape of Banking Malware
Desk Research
Phase 2: Quantitative data collection and metrics identification
How is bank-size delineated in current literature and how can bank-size be measured?
3
Bank-size Metric
4
Data Collection
5
Data Preparation
6
Define Target Frequency Metric
Phase 3: Quantitative analysis
Do bank-size, language-use on bank' website, domain popularity, brand value, being part of banking group and two-factor authentication influence whether a bank is targeted or the intensity of the targets?
6
Descriptive Analysis
7
Regression Model and Results
Statistical Data Analyse
Phase 4: Qualitative data collection and analysis
Which (analysed) banks' characteristics are essential in target selection according to experts?
8
Expert Interviews
8
Analysis of Interviews
Expert Validation
Interpertation
9
Discussion and Limitations
10
Conclusions and Recommendations
11
Future Research
12
Scientific and Social Relevance

Figure 1.2: Research flow

- Chapter 2 provides more extensive information on the use and user of banking malware, namely the threat landscape of banking malware.
- Chapter 3 reviews bank-size measures in the current literature and defines which metrics should be used in this research.
- Chapter 4 explains the external data search process including the limitations in the use of data.
- Chapter 5 explains the general specifications of the Fox-IT data set and commences with the descriptive analysis.
- Chapter 6 prepares the data for the regression analyse.
- Chapter 7 establishes the regression model and discussed the result.
- Chapter 8 describes the lessons learned from the expert interviews and validates the result.
- Chapter 9 discusses the research and describes its limitations.
- Chapter 10 provides an answer to the research question and proposes some recommendations.
- Chapter 11 identifies the next steps for future research.
- Chapter 12 closes with explaining the societal and scientific relevance of this research.

Summary - Chapter 1

The European financial sector is a frequent victim of banking malware leading to massive losses. It appears that banking malware attacks have a narrow geographical focus. There is a limited understanding of why certain financial institutions are more attractive to cybercriminals than others. Understanding the underlying reasons and trends behind banking malware attacks can support decision-making about technical and policy-based mitigation and may even reduce the effects of banking malware attacks. For this reason, this research aims to answer the following question - **Which characteristics of European banks influence the target selection process of banking malware?** By extracting information from configuration files, this research analyses if bank-size, two-factor authentication, domain-popularity and the language-use on a banks websites are indicators that determine if and how often a bank will be targeted.

The next chapter explores the threat landscape of banking malware. It also explains what is meant by ’being targeted’ by describing the technical victimisation process of banking malware.

2
The banking malware landscape

As specified in Chapter 1, banking malware is a class of financial malware. Financial malware is a specialised type of malicious software designed to enable fraudulent transactions (IBM Software, 2014). Traditionally, financial malware is utilised in two manners. The first is an attack against the infrastructure of the financial institutions itself where attacks attempt to transfer large sums in fraudulent inter-bank transactions (Wueest, 2017). The approach against the banks' infrastructure requires a tailored and sophisticated attack to bypass the banks' security measures. This type of attack belongs to the group of Advanced Persistent Threats (APT). APT's consist of a broader range of attack vectors and is not only interested in financial gains. The second approach relates to committing financial fraud or theft by gaining access to a customer's bank account followed by an unauthorised transfer of money (Financial Fraud Action UK, 2018). This approach, using banking malware or Banking Trojans, focuses on financial institutions and aims to target as many people as possible. This research focuses on the second approach; the banking malware attacks.

This chapter describes the banking malware threat landscape by explaining which actors are active in deploying banking malware, how malware victimisation works, and which techniques are being used. In particular, the process of victimisation is essential for this research and for the modelling part later in this thesis. The core concepts of the STIX structure are used to describe the threat landscape. STIX is short for Structured Threat Information eXpression and is a language to convey the full range of cyber threat information in a structured manner (Barnum, 2014). The following three core concepts support describing the threat landscape: *threat actors, common Tactics Techniques and Procedures (TTP)* and *banking malware campaigns.*

2.1. Threat actors

A threat actor is a person/group/organisation who initiates the attacks, either with malice or by accident, taking advantage of vulnerabilities to create loss (Rosenquist, 2009). This section describes the malicious actors; the cybercriminal network and the money mules who funnel out money by forwarding transfers through their personal bank accounts.

Malicious actors

The cyber actor typology of the Netherlands National Cyber Security Centre is used to distinguish and classify the malicious threat actors in the financial malware threat landscape in an organised manner (de Bruijne, van Eeten, Gañán, & Pieters, 2017).

The general motivation behind the Banking Trojans actors is financial gain. In the framework of de Bruijne et al. (2017), this is specified as economically motivated. The framework mentions extortionists, information brokers, insiders, crime facilitators, digital rob-

Table 2.1: Overview of threat agents in the financial malware threat landscape

Threat actor	Description	Expertise	Resources
Extortionists	The extortionist conducts fast growing, widespread and largely not-focused attacks with ransomware. Everyone could become a target: citizens, enterprises, hospitals, schools or even governments. This threat actor can work in a hierarchy, market or network and is personally or economically driven.	low-medium	low-medium
Information Brokers	An information broker is a threat actor who trades information with other criminals like credit card information. The expertise is defined as medium to high since information is unique and valuable data and zero-day information requires some expertise.	medium-high	medium-high
Insiders	Insiders are internal actors who target an organisation from the inside out. These are mostly long-serving employees that are personal, economic or ideological motivated.	medium-high	low
Crime Facilitators	Crime facilitators provide the technical support to the attacks of other criminal actors; e.g. the Dyre-malware campaign. The facilitation could be renting of botnets, development of Remote Access Tools (RATs) or exploitation of kits. These are labour intensive, and facilitators are markets and networks.	medium-high	medium
Digital Robbers	The Digital robbers target financial services, citizens and other enterprises and are economically motivated. Some of the attacks use a sophisticated and long kill chain and require months of preparation and execution.	medium-high	medium-high
Scammers and Fraudsters	Scammers and fraudsters use social engineering in their attacks on targets, such as citizens, enterprises and the public sector.	low-medium	low-medium
State Actors	State-actors target enterprises, public sector or critical infrastructures to gain access to strategic information. They conduct secretive attacks for espionage purposes and are geopolitical motivated.	medium-high	medium-high
State-sponsored networks	State-sponsored network are state-affiliated groups organised in networks. They target citizens, enterprises, public sector, and critical infrastructure and are ideological motivated.	medium-high	medium-high

bers, scammers/fraudsters, state actors and state-sponsored actors as actors who are both economically motivated as well as wanting to target enterprises. State-sponsored actors are designed by a state to use espionage, to steal intellectual property, to hack financial data from banks, and to interrupt operations (Khan, 2018). The criminal ecosystem of banking malware mainly exist of crime facilitators, digital robbers and state-actors, operating individually or in groups.

Money Mules

Money mules are means for threat actors to cash out their illicit funds (Mikhaylov & Frank, 2017). In some cases, the mules are aware of their illegal activities, and others are being tricked into it. A job scam is one example where people are tricked by believing they apply for a job at an innovative start-up, but instead, become a money mule and use their bank account to transfer illegal funds. The idea behind money mules is that the stolen money is transferred to the money-mule account (inside the bank) which is traceable and reversible. Then, the mules send the money to the cybercriminals making the transaction untraceable and irreversible (Mikhaylov & Frank, 2017). Criminals make use of vulnerable people such as drugs or gambling addicts to recruit money mules. Students and high school boys also appear to be easy recruitment targets as money mules. They lend the criminals their debit card and bank account for a certain amount of money (Tajalizadehkhoob, 2013).

Networks

The criminal ecosystem behind the banking malware attack exists of a variety of individuals and criminal groups. Due to the sophisticated nature of the attacks, a network of highly skilled experts is needed to fulfil the various activities. The networks can operate within an organised cybercrime group, or collaborate via purchasing/selling services form the underground market. On these forums services along the attack chain can be bought. Even organised cybercrime groups make use of the underground forums to purchase exploit kits or for money mule recruiting services when they lack this expertise. Moreover, even the entire source-code of malware, kit malware, can be bought or it can be rented paying a monthly fee to operate malware (malware-as-a-service or rented malware). These services make it highly appealing for low-skilled criminals to execute banking malware attacks.

The typical roles of a network comprise administrators, experts, vendors and mules. Figure 2.1 visualises the skills and proportion of the actors operating in the criminal ecosystem of banking malware. The administrators are the moderators of the underground forums. Experts could, among others, refer to malware writers, intrusion specialists and money specialists (Centre, 2017). Malware writers write and update the malware code. The intrusion specialist installs malware on the victims' computer and ensures the malware presence. In the case of Dridex, they may spend weeks and months navigating the network to find the precise machines that they need to access to initiate and validate a large payment. The money specialist is responsible for monetising compromised bots and taking care of fraudulent transactions (Europol, 2017). These are just a few examples of expertise along the banking malware chain, but it gives insight into the complexity of the operation and the need to work in a network.

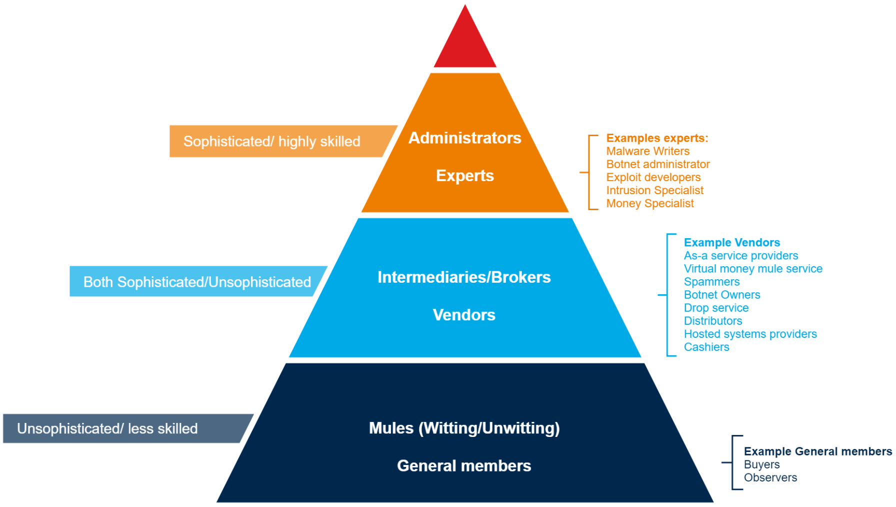


Figure 2.1: Proportion of participation in banking malware attacks

2.2. Tactics, Techniques and Procedures

This section explains the TTP 's of banking malware attacks. To understand how banking malware infects victims and how it can harvest credential, knowledge of the malware components is required. At the end of this section, a typical Banking Trojan attack is explained and visualised.

Architecture of financial malware

Financial malware consists of the following components: the malicious executable, configuration files and functionality enablers (IBM Software, 2014).

(A) The **malicious executable** installs the malicious software on a device. This dropper, for example, an *.exe* file, enables the malware to continue running even after attempts to remove it from the computer or when a system is rebooted.

(B) The **configuration files** is a binary-format file (block of computer memory) that determines the targeted application and the malware behaviour in each of these applications. The man-in-the-browser attack vector is an example, where the following elements define the behaviour of a typical man-in-the-browser attack:Command and Control (C&C) server information (IP or server name), web injection, back-connect configuration, web filters, Domain Name System (DNS) redirect configuration and grabs.

Web injection list all the URLs (could be thousands of records) where content should be injected or modified. In other words, it contains all the domains that the malware writer wants to target. *Back-connect configuration* ensures that it looks as if the traffic originates from the victim's IP. *Web filters* are developed to prevent communication from the machine to the anti-virus management servers, disable the device to update or prevent the browser from logging out. The *DNS redirect configuration* blocks a specific site and redirect it to another domain name with an invalid IP address. *Grabs* gathers information from a web page by a submitted HTML (by the victim) or by taking a screenshot.

Subsequently, the configuration file has two parts; the static and the dynamic configuration. The static configuration is a piece of code, containing information of the URL, the IP address or host name of the C&C server, the customer identifier and the encryption key. This information is required at the moment that the bot is executed.

The dynamic configuration is downloaded by the bot from the C&C server of the attacker. It contains information on websites (to target or to ignore) but also about the action to execute when a specific website is visited. The dynamic configuration gives the attacker greater flexibility in controlling the attack; by changing the targeting requirements or by attempting to bypass a banks security measurements (Black, Gondal, & Layton, 2017). The dynamic configuration enables the adversary to interact.

(C) The **functionality enablers** are software components that execute a desired functionality. For example, legitimate software used for malicious purposes, or specific actions developed by a fraudster such as screen-capturing software or key-loggers (they monitor each keystroke typed on the victims' computer keyboard). They can be built-in or downloaded on demand from the C&C server.

Command and Control server

The Command and Control server components communicate with the infected client to send information or to retrieve updates. Among the capabilities of the C&C server are: reporting on the number of infected machines (botnets), distribute commands to specific bots or all bots, distribute configuration file updates or software upgrades and manage the harvested credentials.

There are three Command and Control architecture forms (Black et al., 2017).

- A botnet can be managed by one or two central servers. The information flows from the bots to the central servers. This type of communication has a centralised architecture, based on the standard IRC and HTTP protocols (Gardiner, Cova, & Nagaraja, 2014).
- Peer-to-Peer (P2P) means that the bots/peers are interconnected nodes that share resources and information among each other; this is a more decentralised architecture. The communication flows between the bots to the threat actor to obfuscate the machine and manage the botnet.

- Domain Generation Algorithm (DGA) structures are used to generate a large number of domain names periodically. Rather than developing a new malware version and to set up the whole cycle again on a new server, the malware automatically switches to a new domain at regular intervals.

Infection Vectors

Infection vectors are methods through which the adversary is trying to distribute the malware to a maximum number of people. The following methods are common for distributing malware: e-mail spam, malvertising, infection services and exploits (Custers, Pool, & Cornelisse, 2018; IBM Software, 2014; Nagunwa, 2014).

- The **email spam** is a well-known distribution vector. People are persuaded via spam to click on links or open attachments, such as a word or excel document, and get infected by Microsoft macros (Custers et al., 2018).
- With **malvertising**, the adversary pays for advertisement space and plants malicious ads with embedded JavaScript. The script uses an existing exploit (vulnerabilities) or redirects the user to the exploit kit to install the financial malware. Besides, via flash player, the adversary can automatically run a script which downloads the malicious content on the victims' computer.
- Adversaries developed a specific type of malware that can distribute malware, such as loaders and botnets. Malware campaigns sometimes use affected machines with old malware to distribute new malware. Anyone can easily purchase this **infection services**.
- The last distribution factor is the **exploits (exploit kits)**. Security vulnerabilities in the browsers or desktop software are ”exploited”. A piece of (payload) is introduced which installs itself in the desired malware.

Harvest credentials techniques

The adversary can use a variety of malware features to harvest bank credentials such as screen captures, keylogging, session hijacking, pharming, web injections, Man-in-the-Middle (MitM), Man-in-the-Browser (MitB), overlay attacks/fake forms, remote-control tools, loaders and rootkits (Custers et al., 2018). A short explanation of these features will follow in the next paragraphs.

- **Key logging** is a piece of software that records users' keystrokes and then sends this information to the C&C server to bypass two-factor authentications (IBM Software, 2014). The adversary steals the authentication cookies and impersonates the users.
- **Pharming** is the hijacking of a domain name service to redirect the end user's browser to a malicious site pretending it is a legitimate banking site. Pharming modifies the DNS configuration of targeted applications, either by changing the IP addresses of specific sites locally or by exploiting a DNS server's vulnerability. A relatively new form of pharming changes the DNS configuration of the targeted application in a local router. In this case, the router redirects the end user's browser to the malicious site. This method has the added benefit of affecting all machines that use this router instead of infecting a single computer (IBM Software, 2014; Nagunwa, 2014; Ståhlberg & Corporation, 2007).
- **MitB** injects fraudulent pages or form fields, which is also called ’form grabbing’. When an infected victim wants to log in on his/her bank account website (defined in the configuration file), the adversary injects a malicious webpage. After the victim inserts its credentials, the credentials will be sent to the C&C server (Ståhlberg & Corporation, 2007). The malware needs to be compatible with any browser and all versions of it (IBM Software, 2014).

- **MitM** attacks against online banking uses a fraudulent website that alters traffic between the user (interface) and the security layer of the server (Ståhlberg & Corporation, 2007).

- **An Overlay attack** waits until the user visits a specific online banking service, defined in the configuration file, and then launches a fake form or website (Wueest, 2017). When the credentials are submitted, the information is sent to the C&C server. Most overlay attacks launch the fake form on top of a legitimate website when the end user accesses a targeted application. The involvement of the end user's browser is required, which differs in this sense from hijacking and pharming attacks (IBM Software, 2014).

- **The Remote-control tool** software enables the adversary to gain full access to an infected device, local files and the end user's browser. It takes over the administrative control via a backdoor. This malware is known under the name; **Remote Access Trojan**. This type of attack is executed from the customer's machine and evades device security measures (IBM Software, 2014).

- **Loaders** are very similar to Remote Access Trojans, but they have only basic capabilities (DLexec, uninstall, DLupdate). These loaders are easier to crypt since they have fewer codes and will remain longer undetectable.

- **Root kit** originally is a UNIX term. 'Root' is the system-take-over-control of the system resources. 'Kits' is the software developed to take root/administrator control of a PC (Cucu, 2017). Rootkits are designed to access a computer unauthorised.

How does a typical banking malware attack work?

Commonly used methods of inserting infection banking Trojans rely on web-based techniques using MitB, key-logger and form grabbing techniques. In Figure 2.2 a typical financial malware attack is visualised.

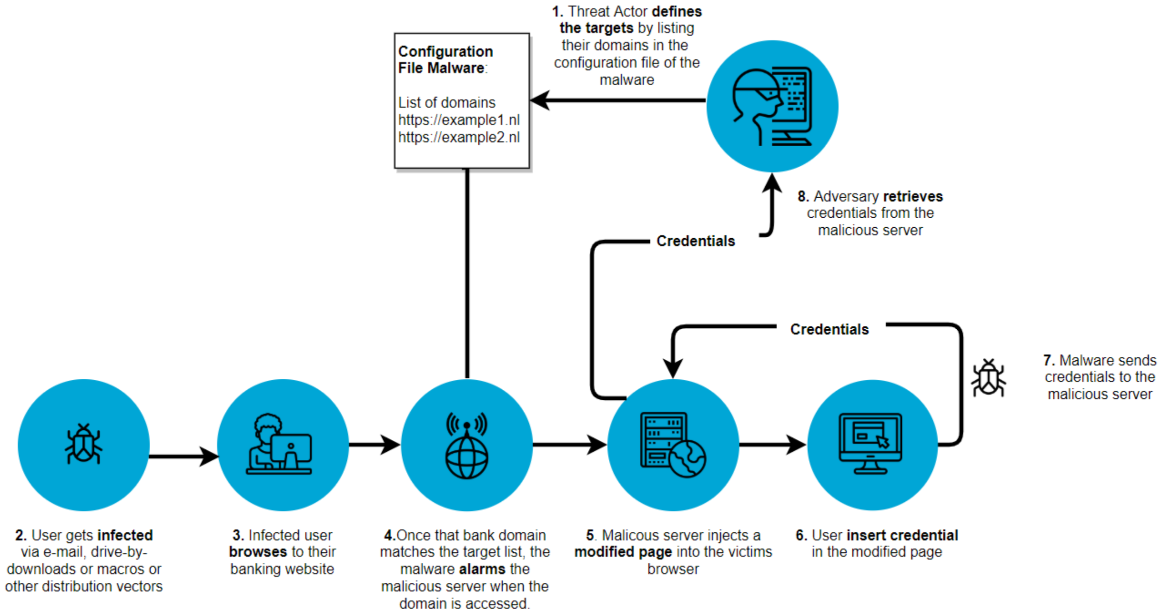

1. Threat Actor **defines the targets** by listing their domains in the configuration file of the malware
Configuration File Malware:
List of domains
https://example1.nl
https://example2.nl
8. Adversary **retrieves** credentials from the malicious server
Credentials
Credentials
7. Malware sends credentials to the malicious server
2. User gets **infected** via e-mail, drive-by-downloads or macros or other distribution vectors
3. Infected user **browses** to their banking website
4.Once that bank domain matches the target list, the malware **alarms** the malicious server when the domain is accessed.
5. Malicous server injects a **modified page** into the victims browser
6. User **insert credential** in the modified page

Figure 2.2: Visualisation of a typical banking malware attack.

2.3. Campaigns: Banking Trojan families

A campaign is a grouping of adversarial behaviours that describes a set of malicious activities or attacks that occur over a period against a specific set of targets (Barnum, 2014). Even though new malware families entered the threat landscape in 2018, also the old Banking Trojans remained active. The Banking Trojans Zeus-Panda and Citadel remained active over the last few years. These Banking Trojans are described in other researches (Charalambous, 2018; Natalius, 2018; Tajalizadehkhoob, 2013) with similar research questions, and therefore there is no need for repetition. Appendix A describes all malware families and campaigns that are active in today's threat landscape and represented in the Fox-IT database. This section provides a short description of Banking Trojans that is not yet covered in existing literature; Gootkit, TheTrick and BokBot.

Gootkit

The Gootkit Trojan steals confidential information and downloads additional files on the compromised computer through spammed email messages. Other files may also be downloaded or delivered silently through web exploits. Gootkit malware is rented to a specific number of people. It has a format similar to ZeuSv2 and operates in Europe and Australia. It targets Windows and has common featured as well as other Banking Trojans: form grabbing, File Transfer Protocol (FTP) /mail stealer, spyware and also video recording (ThreatConnect, 2016).

The features of Gootkit exist of Virtual Networking Computer (VNC), video capture, webpage injection, process injection and a sock tunnel. VNC enables the adversary to access the victims' computer remotely. The SOCK is a SSH-based proxy on the device and enables port forwarding. The purpose of this is to establish a two-way communication tunnel with the device, which allows the attackers to use the victim's IP when accessing the compromised bank account.

TheTrick

TheTrick or TrickBot has been discovered in 2016 and is ever since continually upgrading its features. It has the overlapping attack module and C&C communication as the malware Dyre (see Appendix A). The malware targets small financial services but is wide-spread across the world. Sometimes the configuration file was updated several times a day. It targets mainly Windows and is a P2P botnet. The primary infection vector is spam, but incidentally, an exploit kit or third party botnet are used (Duncan, 2018). Trickbot prompts a screen lock to ensure credential input. Trickbot abuses the Eternal Blue vulnerability exploit. These are vulnerabilities in Microsoft's implementation of the Service Message Block (SMB) protocol. The vulnerability of the first version of SMB allows malware to move laterally to infect connected devices.

Since June 2018, Trickbot is moving from an infected Windows client to a vulnerable Active Directory domain controller. Trickbot's lateral movement over SMB is distinctly different from WannaCry's implementation of EternalBlue noted in 2017, so this method of SMB propagation appears to be based on a different exploit developed by Trickbot authors (Duncan, 2018). Trickbot has its malspam-based distribution channel, but recently Trickbot attackers are also using Emotet for their infections.

BokBot

BokBot came to the attention of security researches at the end of May 2017. It is a combination of Zeus GameOver, Citadel and Corebot (Klason, 2018). The geographical focus of BokBot is mainly the US, but it has also spread to other countries to gather credentials for online services.

BokBot supports three different types of configurations which are all in a binary format rather than in structured formats like XML, as used by TheTrick (Klason, 2018). The other two configuration are used to control how the bot will interact with the targeted entities such as redirecting and modifying web traffic related to, e.g. internet banking for harvesting credentials and account information.

The reporting configuration is used for a universal purpose, where it is not only used for *screenshots but also for HTML grabbing* which would grab a complete HTML page if a victim browses to an “interesting” website, or if the page contains a specific keyword (Klason, 2018). This information enables the actors to conduct some reconnaissance for future attacks, like being able to write web injects for a previously unknown target.

Summary - Chapter 2

Banking malware is aimed to commit financial fraud or theft by gaining access to a customer’s bank account. Bank customers, using online bank services, can be targeted by adversaries who have an array of technical software tools at their disposal to infect and harvest their credentials. Due to the sophistication of banking malware attacks, cybercriminals never act alone but operate in a broader network or buy services along the attack chain from underground markets.

A typical banking malware scheme starts with the threat actor listing the domains of banks in the configuration file. When the infected user browses to the banking website (which matches with one of the domains listed in the configuration file), the malware orders the malicious server to inject a modified page. The user inserts the credentials in the modified page, and the information is sent to the command and control server of the adversary.

The next chapter takes the first step to explore the relationship between bank-size and target selection (Research objective in Chapter 1). It selects a suitable bank-size metric that fits this research best.

3

Measuring bank-size

Chapter 2 is a necessary step to recognise the wide range of options adversaries have to interfere in online banking activities when using sophisticated software. By describing the technical implementation process of banking malware, the chapter explains what is meant by whether a bank is being targeted. This understanding of the cyber threat landscape and the target process is essential as preparation for the analysis.

The objective described in Chapter 1, is to seek a relation between bank-size and target selection. This chapter will further analyse this relationship. Various bank-size measurements used in academic fields are listed and analysed in order to ultimately decide upon one suitable bank-size metric. Not all bank-size measurements used in academic literature can be used for this study. This chapter aims to find the most suitable bank-size measures that fits this research best.

3.1. Traditional bank-size metrics

Multiple researchers define the systemic risk and organisation level in relation to bank-size measures. This section will evaluate how these studies define bank-size.

Systemic risk

Systemic risk is the systematic failure of multiple, disparate systems due to a single event (Romanosky, Ablon, Kuehn, & Jones, 2017). Academia analysed if bank-size measures increase systemic risk, thus if the failure of large banks has a more disruptive effect on the financial systems as a whole. Laeven and Ratnovski (2014) provide strong evidence that systemic risk increases with bank-size when bank-size is measured as the natural logarithm of total assets. Also, Chen, Office, Damar, and Soubra (2012) provide evidence that the marginal contributions of individual banks to the systemic risk indicator are determined mostly by bank-size. Here, the institutional size is measured by the total liability. Furthermore, Arinaminpathy, Kapadia, and May (2012) explored the role of large banks in systemic risk. They defined bank-size and interconnectivity as metrics to reflect the systemic importance. Here, bank-size is not measured in absolute terms but as a proportion of the system's initial total assets.

Organisation risk and performance measurement

The management of risk, assets, and liabilities are functions of banking. There are multiple risk methods that define the risk of a financial service or measure its performance. These assessments are vital since the volatility of liquidity risks can give an early signal for a banking crisis. Waemustafa and Sukri (2016) investigated the influence of external and internal factors affecting liquidity risk of Islamic and conventional banks. Here, bank-size was defined as the natural logarithm of total assets of a bank in years. Furthermore, the liquidity risk was calculated by multiple factors that were standardised on *total assets, total equity* and *total loans*. These common variables also came back in the research of Lepetit, Nys, Rous, and Tarazi (2008) and Jones (2005).

Baron (2015) defined his metrics for banking performance. He uses the following variables: net income, shareholders equity, total assets, net interest income, average earning assets, non-interest expenses, net revenue, total loans, total deposits, capital, Risk-Weighted Assets and book value per share. The research of PWC (2011) used also net income as a variable.

Firm-size measurements

Dang and Li (2015) assessed the sensitivity of empirical results in corporate finance to different measures of firm-size. This research employs three popular firm-size measures to 100 empirical studies; total assets, sales and market value of equity. This study shows that firm-size measures are robust and statistically significant.

European measurements

The European Bank Authority (EBA) defines bank-size of EU banks by the assets as a percentage of total consolidated assets. The banks are ranked across three groups: large banks account for more than 0.5%, medium between 0.5 and 0.005%, and small banks for less than 0.005%. These standards are also used by the European Parliament (de Groen, 2016). Another example of a used standard is when EBA conducted a risk assessment to assess the quality of credit ratings assigned to banks in Europe and in the United States. Here, bank-size was presented by the natural log of the book value or, in other words, the *shareholders equity* (Marques-ibanez, 2012).

Commercial measurements

Forbes Global 2000 uses four measures: assets, sales, profits and market capitalisation. Fortune500 uses sales and profits. They both use sales and profits, which seldom appeared earlier in academic research.

Concluding, the following bank-size measurements are used in traditional economical researches research, in the European Union and Private Sector:
Traditional bank-size measures: Total assets, inter connectives, total liabilities, total equity, total loans, shareholders equity, net income and return assets.
European standards: assets as a percentage of consolidated total assets, shareholder equity.
Commercial standards: assets, sales, profits, and market capitalisation.

3.2. Bank-size metrics in the cyber domain

Bank-size indicators, such as capital and loans, are rarely considered in the cyber risk assessment. This chapter explores the field of cyber risk analysis and cyber insurance to determine how bank-size is represented in current literature.

Cyber risk analysis

Risk in the cyber domain is often defined as the impact/loss times the frequency/probability of occurrence. Losses can be in terms of assets, reputation or organisation (Jones, 2005). Logically, large institutions face more substantial losses when an attack occurs. However, the quantitative assessment such as FAIR (Jones, 2005), did not have a standardised calculation for these losses. The potential impact or losses of an organisation are defined by the risk analyst, the organisation or even an insurance company (Lund, Solhaug, & Stølen, 2010). Thus, while bank-size indicators are included, the risk level is based on human interpretation.

The size of the firm can be used for the calculation of potential losses, since they are related to each other. For example, a data-breach in a firm with a high number of clients will lead to a higher damage value compared to a firm with fewer clients. However, this is rarely considered in quantitative assessments in the literature. Only Custers et al. (2018) set-up a formula on productivity loss. He defined this by multiplying employees, the hours of downtime and the average wage. Calculating productivity loss is one of the examples where the *number of employees*, is used as bank-size indicator to calculate loss (part of risk).

Cyber insurance

Cyber insurance is an insurance product used to protect businesses and individual users from Internet-based risks, and more generally from risks related to the information technology infrastructure. Cyber insurance premiums could depend on many factors: the company's industry, services, type of sensitive data, the total number of Personal Identifiable Information (PII), Personal Health Information (PHI), data risk and exposures, computer and network security, privacy policies/procedures and annual gross revenues. Romanosky et al. (2017) discovered that many policies used a very simple metric to price their premiums, namely factoring it on the *expected loss* of an attack or data breach. It is unknown how expected loss is measured since there are no standard metrics.

A standard bank-size metric for the cyber domain is not represented in current literature. Though, the number of employees or number of clients associated with the expected losses are used to make calculations.

3.3. Bank-size metrics in target selection studies

Multiple experts and researchers mention that larger organisations are more prone to be targeted than smaller organisations; the size of an organisation seems to be an important factor.

Table 3.1 provides an overview of all target-selection studies investigating if organisation size correlates with the number of targets. The table shows that the number of clients and employees are popular measurements. Cheung (2017) also introduced financial and market characteristics, such as profit, market value and net income. These studies used the Fortune500 websites as dataset and only analysed banks listed in the Fortune500, which might result in a biased conclusion.

The next paragraph will list bank-size measure indicators. For each of them one or more advantages and disadvantages will be mentioned. Some of these indicators are also listed in Table 3.1.

3.4. Selecting bank-size measurements

Not all bank-size measurements, available in academic literature, can and need to be listed and analysed. This section focuses on the most frequently mentioned bank-size measurements in well known studies. It describes also some of the pro's and con's of the selected bank-size measuring indicator. Some of the indicators will be suitable for the modelling part later-on in this study.

Revenues

Revenues are the total income, also named as the gross earnings. It is the sum of net interest income, fee and commission income, trading income and other income. Revenues indicate what customers are prepared to pay for the provision of a particular service. Schildbach and Schneider (2017) state that revenues are the most comprehensive, comparable and robust indicators. He recommends to use multiple indicators for bank-size, but among them revenue seems to be the best one.

Advantages

- Relatively stable and less volatile than market indicators, such as market capital. It represents the core operating trend better than the stock market.

Table 3.1: Overview of target selection studies analysing organisation size

Organisation Size	Attack Type	Research Outcome
Employees	DDoS	Organisations with more than 500 employees are likely to experience a DDoS attack, incur higher attack costs, and require more employees to mitigate the threat (Arbor Networks, 2016).
Profit, Market Value, Net Income	DDoS	The quantitative analysis of Cheung (2017) showed a weakly correlation for organisation size; only market value is considered an important indicator for the number of attacks. The question is raised if organisation size indeed relates to target selection of DDoS attacks or if there is an inherently related factor, such as media attention.
Customer Base and Wealth (Deposits)	Zeus Malware	It is observed that size is a threshold for US banks getting targeted by Zeus Malware, but this did not predict the intensity of attack (Tajalizadehkhoob et al., 2014).
The number of clients	Zeus Malware	Ten SEPA countries. According to the quantitative analysis, between 2009 and 2013q1, financial institutions with more clients encountered a higher attack intensity. However, this only accounted for three out of the ten countries and is only significant for the year 2009. For the first nine weeks of 2013, financial institutions with fewer clients are targeted more frequently.
Number of Employees	Spear-Phishing Attacks	Symantec (2015) found that both large and small business appear to indiscriminate. For the year 2011 till 2015 the analysis has shown that businesses with less than 250 employees were targeted more often.
Revenues	Malware	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.

- It is based on realised numbers and not expectations, such as market capital.
- It doesn't rely on complex valuation models.
- It is a comprehensive all-encompassing measure of the banks total operations. Banks also play roles as originators, services and intermediaries and can earn substantial fees and commissions. These are represented in total revenues, but not in total assets.
- It reduces banks activities to a common denominator where only cash flows are taken into account.

Disadvantages

- It can be inflated by non-bank activities, such as insurance, without affecting the balance sheet. Since European banks mostly are focused on core banking, it will not play an important role in this analysis.
- In some cases it could also be a measure of risk instead of size. High risk banking activities may lead to high revenues but the total of revenues might still be very limited in the total financial system of banks.

Total Equity

Total equity measures the value of a bank. It does not reflect the banks business as good as revenue, but it is still a reasonable estimate. This indicator is similar to total assets. It looks at the value of the bank instead of the volume of the transactions. Total equity is a book value that can be found in the balance sheet of the bank. It provides a more neutral view on the size irrespective of whether it is generating none, little or a lot of money with its activities. There are several ways to measure equity: shareholders equity, tangible equity, total equity and Tier 1 capital and common equity. This research uses *total equity* since this it is a stable, independent, common available and easy accessible measurement.

Advantages

- It doesn't contain hypothetical calculations.
- It is independent from differences in business models or financial system structures.
- It is a stable measurements over time with limited fluctuations.

Disadvantages

- Equity is subjected to changes in accounting rules. For example, some losses are not recognised by Profit and Losses.
- The indicator is also related to historical values which do not reflect the current value.

Total assets
Total assets is the sum of the total on the balance sheet of a bank. The balance sheet total indicates the gross volume of all combined exposures and is not weighted by risk from loans, security holding and derivates.

It is the most commonly known representation in academics and is often used as a threshold value to categorise banks as small, medium, or large. Even though total assets are frequently used in literature, it is not the best measurement according to Schildbach and Schneider (2017) because it suffers from significant valuation problems and it does not account for differences in individual bank business models and financial systems.

Advantage

- It is the most prominent size indicator used by official representatives and relatively easy accessible in external sources.

Disadvantages

- Loans are part of total assets. They are sometimes not repaid, and therefore the banks must predict the risk, which is difficult to do.
- Derivatives or credit default swaps, as part of total assets, can be measured differently (depending if EU or USA regulations are applied).
- It is not clear which metrics are used in the definition.

Market capitalisation
Market capitalisation is the market value of the company's shares, or in other words the total value of all of a company's outstanding stock. It is related to the stock prices and takes different aspects into account that don't appear in the balance sheet.

Advantage

- It is open-source information and also criminals might use it. According to the expert interview of Natalius (2018), the relative market share of a bank could be a potential measurement.

Disadvantages

- This measure may show many and strong fluctuations.
- Dependent on stock market sentiments.
- No clear view on income through service/ market activities.

Net income

Net income is defined as sales minus expenses (such as taxes interest and others). Net income is an input factor for market capitalisation. It is an indication of how profitable the company is and it is also used for performance measurements.

Advantage

- It is accessible open source information.

Disadvantages

- It only provides a partial view on bank size.
- May show big fluctuation in time.

Risk-Weighted Assets

Risk-weighted assets account for the minimum amount of capital that must be in place for an institution to reduce the risk of insolvency. It is an indicator which is only found in the banking industry. In the financial crisis of 2007 and 2008 the banks lost a significant amount of their capital for an array of reasons. One of the reasons is that many consumers failed to pay the interest of their mortgages. To avoid this happening again, regulations determine that banks group their assets by risk. The Risk-Weighted Assets (RWA) help preventing banks from losing their capital which could then sharply decrease the value.

In Europe, this regulation is the Minimum Requirement for own funds and Eligible Liabilities. It ensures that the banks have sufficient capital and debt instruments available for a bail-in (rescue of a financial institution that is on the brink of failure whereby creditors and depositor take a loss on their holdings).

The research of de Groen (2016) state that smaller banks (based on the EBA standards) have a higher average risk weight and that medium-size banks have almost a quarter higher average risk weight than large banks. It is unsure if RWA is a good measurement to represent size.

Advantages

- It is a good and broadly accepted indicator.
- It counts for 42 per cent of the total assets (de Groen, 2016), but RWA differs substantially from total assets. It counts for different parts of the asset side of a banks balance sheets and takes recognised risks into account.

Disadvantages

- It depends on internal models, accounting rules and regulations (Europe differs from the US).
- Earlier research has shown doubts about the reliability of the measure.
- Risk weights are estimations.

(Online) Customers

Banking malware targets and tries to victimise customers from a certain bank and therefore cyber criminals will target as many online banking customers as possible and will thus try to target banks with a large number of customers (Leukfeldt & Yar, 2016). The number of customers proved in multiple target selection studies to be an essential factor for cyber activities.

Advantages

- It is a proven factor for online banking and cyber crime activities (see Table 3.1).

- Easy accessible open source information.

Disadvantages

- It is not really reliable: many clients could still generate relatively small amounts of money.
- Earlier research didn't prove a stable and reliable relation between number of clients and number of attacks.

Number of employees

The number of employees appeared as a measurement for bank-size in empirical work (Dang & Li, 2015) and is a researched indicator in cyber-related studies (Cheung, 2017; Symantec, 2015). A possible assumption underlying the use of the number of employees in cyber-related studies is that they indirectly represent capable guardians to prevent attacks.

Advantage

- It has been subject of research earlier (Table 3.1).

Disadvantages

- Similar to (online) customers; retail banking is quite labour-intensive compared to asset management.
- An inefficient banking organisation may have many employees but may still show a low profit. This measure would probably not give the best indicator on an individual level, but together with more profit related indicators, it could work.

Branches

According to Schildbach and Schneider (2017) this is a clear measure with similar business models and geographic orientation. A bank with branches all over Europe implies that it is a successful bank, which could be associated with value.

Advantage

- It has been subject of earlier research; structure shows the magnitude of the organisation.

Disadvantage

- The drawback of using branches is that the branch density is very different in countries like Spain, France and Italy compared to the Netherlands and the UK.

Deposit of customers

Deposits of customers are introduced to deal with the drawback of using (online) customers as a metric. This is because fewer clients could have on average more money on their bank account. Since the people behind banking malware are financially driven, they would probably choose high amount of deposits over the number of customers as a decision-making factor rather than the number of customers.

Advantage

- Banking malware aims to infect customers in order to gain access to their deposits and is, therefore, a suitable measure.

Disadvantage

- It is difficult to obtain information on the deposits of customers without searching through the financial reports. For this reason, cybercriminals will probably not investigate in the number of customer deposits as part of the reconnaissance phase.

Loans of customers

A considerable percentage of bank assets are loans. The factor of the total amount of loans is often used in the financial world as traditional bank-size measures.

Advantage

- It has been subject of earlier researches.

Disadvantage

- Banking Trojans aim at victimisation of the deposit of bank' customers. Loans is not a robust representative indicator for an increase target frequency of a particular bank.

Based on the Schildbach and Schneider (2017) ranking, literature, and previous target selection studies the information provided in Table 3.1 the bank-size measuring indicators listed in table 3.2 are selected and will be tested to find out whether they are a good representation of bank-size in the banking Cyber domain.

It should be noted that the chosen measures capture several dimensions of the 'firm-size'. Market capital is more strategically oriented on a highly uncertain future, whereas Total sales are more oriented on the product market and is less strategically forward-looking.

Table 3.2: Selection of bank-size measurements

Indicators measuring bank-size.	**Main rationale of selection**	**Ranking in Schildbach and Schneider (2017) study**
Revenues	Revenue is a common denominator for a wide range of bank activities, are cash flow-based, thus do not depend on business model or financial structures.	No.1
Total equity	Provides a neutral view on pure size, does not depend on hypothetical calculations and is a stable measurement over time.	No.2
Total assets	Most prominent and favourable size indicator according to official-sector representatives. EBA and other European institutions use this indicator to measure bank-size.	No.3
Market capital	Focused on the m arket value which includes different aspects on a balance sheet.	No.3
Net income	Traditional measure for bank-size.	Tile 4: other indicators.
RWA	Represents several parts (focus on the minimum requirement for own funds and eligible liabilities) of the balance sheet aside from total assets.	Tile 4: other indicators
The number of customers	Banking malware aims at victimising customers and it appears in several target selection studies.	Tile 4: other indicators
Number of employees	Valid metrics used in traditional empirical research and cyber-related studies: e.g. as target selection for DDoS.	Tile 4: other indicators
Branches	A measurement with similar business model and geographic orientations, which suites the research purposes of this study.	Tile 4: other indicators
Loans customers	Total loans is a commonly used traditional measure.	NA
Deposit customers	Deals with the drawback of using customers by showing how much money could be a potential gain for the financial malware-deployer.	NA

Summary - Chapter 3

The goal of this research is to add bank-size as financial factor to the explanatory model in order to get a comprehensive view concerning target selection. How bank-size is measured in traditional, cyber and target selection researches is explored to identify suitable measurements for bank-size. A list of bank-size measurements is selected based on Schildbach and Schneider (2017) ranking (Table 3.2), current literature, and earlier target selection studies. These measurements are: Revenues, Equity, Total assets, Market capital, Net income, RWA , Number of (online) customers, Number of employees, Branches, Loans and Deposits of customers.

Next chapter will collect data of the above selected bank-size measurements and other banks' characteristics that is needed for the analysis of this research.

4 Data collection

The aim of this research is to include bank-size in the explanatory model. Chapter 3 completed the first step by selecting suitable bank-size measures that will be analysed. This chapter proceeds by collecting the data needed for the analysis. The entire data collection process: scope, reliability and the limitations of the data, will be extensively described in this chapter.

4.1. Introduction to databases

Data needs to be collected for the bank-size measures and the other characteristics of the bank. Figure 4.1 visualises the inter-connectivity of the databases and shows how the databases will be used and connected. This chapter will also explain how the databases are acquired from the external sources, what their scope is, how reliable the dataset is and the related limitations. The following datasets will be partly used in the analysis.

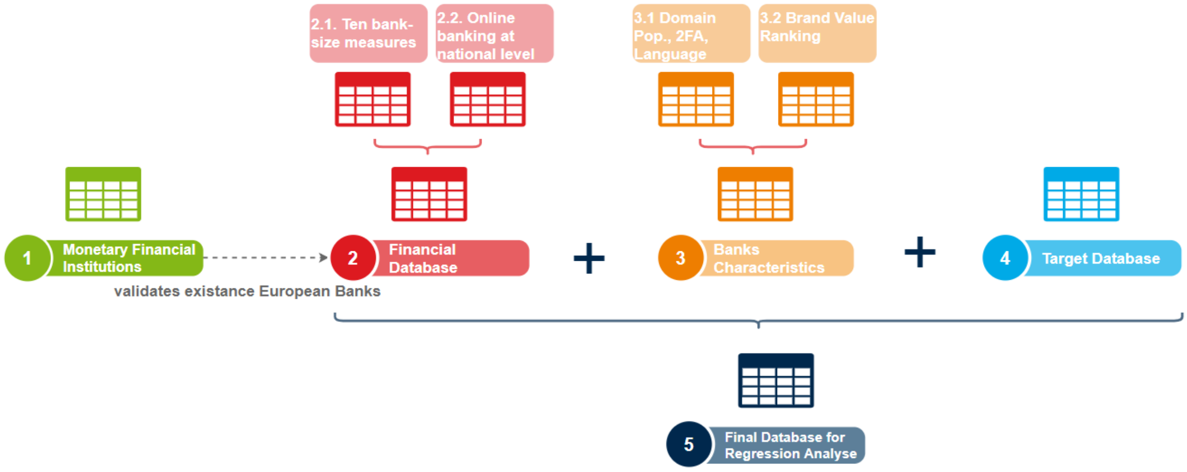

2.1. Ten bank-size measures
2.2. Online banking at national level
3.1 Domain Pop., 2FA, Language
3.2 Brand Value Ranking
1 Monetary Financial Institutions
validates existance European Banks
2 Financial Database
+
3 Banks Characteristics
+
4 Target Database
5 Final Database for Regression Analyse

Figure 4.1: Relation of utilised databases

.

1 **MFI database**
This is a list containing all individual European banks. It is issued by the European Central Bank (2018).
The available variables are: Bank name, Country, Office address and HQ profile.

2 **Financial database**

2.1 **Bank-size measures database**
This contains the available information of the selected bank-size indicators for the European Banks. Information is collected from the bankers-database, the annual financial report and the central banks documentation.
The variables are: Revenues, Total Equity, Total assets, RWA, Net income, Number of customers, Number of employees, Branches Loans of customers and Deposits of customers.

2.2 **Online banking database**
This database is retrieved from (Eurostat, 2018) and contains information on the average percentage of online banking in each European country. This dataset is used to convert the number of customers to online customers.
Variables are: the percentage of the online customers out of the total customers per country.

Appendix B provides an overview of the utilised sources for the financial database on country-level. The source-information per bank can be found in the financial database, in column ”Source”. This database is made publicly available on github: `https://github.com/MarritH/TargetSelection_BankingMalware`.

3 **Target database**
This database provides information on the target frequency on the European banks and is extracted from Fox-IT malware lab.
The variables are: Attack URL, Entity profile (domain, name, country), malware variant, time, being part of a banking group

4 **Banks' Characteristics database**

4.1 **Two-factor authentication, language and domain popularity**
This is the information on the two-factor authentication, domain popularity and language of each bank, acquired from the research of Natalius (2018). This database only provides valid information for online services.
The variables are: the two-factor authentication, the domain popularity, the language, if it is a customer bank and if it provides online services.

4.2 **Brand Value Ranking**
The brand value ranking data is retrieved from (Brandirectory, n.d.). It is not the most reliable data since it is obtained from a website. However, it provides a lot of information on the ranking based on brand value and is easy to obtain. Also, banks from the USA and China are added to this rank but these banks are out of scope and, thus, removed.

The following four sections provide additional information on the scope and limitations of the databases.

4.2. The Monetary Financial Institutions database

The information on European banks is acquired from the EBA (European Central Bank, 2018). It lists the Monetary Financial Institutions (MFI)s who are resident credit institutions and almost by definition member of the EBA (Article 4 of Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013). The database is considered reliable since it is from a legitimate European institute. The information has been downloaded from the official EBA website.

4.2.1. Scope

The EBA distinguishes four types of MFIs: the Central Bank, Money Market Fund, Credit Institutions and Other Institutions. Not all of these institutions are relevant for this research. A Banking Trojan targets individual costumers by pushing or pulling malicious software that will execute its work when customers log-in in their online banking page (see section 2.2). The scope is set by discussing the four types of MFI's on their relevance given that banking malware targets the customers of the banks. Only credit institutions appear to be relevant in researching banking malware.

Central Banks

Central banks have a responsibility to supervise and provide a monetary policy of their respective financial institutions. These institutions work closely together with the Central Banks. Costumers have no direct or individual contact with the Central Banks. For this reason, Central Banks are not considered as part of the analysis.

Money Market Fund

According to the EBA, Money Market Fundis a collective investment undertaking. These funds primarily invest in money market instruments with a residual maturity up to one year or in bank deposits (Article 2 of Regulations ECB/2013/33). Money market fund institutions could have login pages where consumers can log in their business/personal banking page. However, investment institutions are more focused on large businesses. For this reason, money market fund are not included in the scope.

Credit institutions

Credit institutions form by far the largest part of the monetary financial institutions sector in Europe. Credit institutions are businesses from whom to take deposits or other repayable funds. They grant credits for own accounts (point (1) of article 4(1) of Regulation (EU) No 575/2013(CRR)). Credit Unions focus and serve a specific community of members such as teachers or members of the military. This specific community is an important element of the marketing strategy. Credit unions are owned by their members and operate for their benefit. In the EBA list, the following countries listed their credit unions.

- Ireland has 278 registered credit unions which account for 14,3 billion of total assets. The credit unions are not-for-profit institutions and are well known for not keeping up with the latest banking technology. Only twelve percent of credit unions have more than twenty percent members doing some business online and twenty-six percent of the institutions have only two or less percent of the members doing 'something' online such as checking their balance (McKillop, O'Connell, & O'Toole, 2016). The trend is that credit unions are merging to cope with the competition of profit banks, especially on IT. Although credit unions account for 14,3 billion of total assets and are thus a large part of the Ireland Banking system, banks should provide and use considerable online-banking services to fit the scope of this research.
- In Lithuania 56 credit institutions belong to the Lithuanian Central Credit Union. The EBA has listed these credit institutions as individual banks. They count for 2,6 percent of the total assets in Lithuanians financial system (OECD Committee on Financial Markets, 2017).
- The Polish credit unions operate under the single brand Spółdzielcza Kasy Oszczędnosćiowo-Kredytowe (SKOK). Poland's credit unions account for $4.8 billion of aggregated assets and serve 2.2 million customers. Since these credit unions operate under one online service, they have been merged.

Although there is information on the Credit Unions, they will not be part of the scope. Credit unions are not-for-profit institutions, account in most countries for a small percentage of the total assets and there is a risk to use facts and figures of institutions famous for not having online customers. This profile does not match the scope and is, therefore, excluded.

Other deposit-taking corporations/institutions
Other deposits-taking corporations/institutions are financial inter-mediators or electronic money institutions. These intermediates receive deposits from institutional units, grant loans or make investments in the security of their account. Such an institution is a legal person who is authorised to issue money electronically (article 2 of Directive 2009/110/EC). Intermediates are not considered to be part of personal banking and will therefore, not be part of the scope.

4.2.2. Limitation

The EBA updates the information on the MFIs on a daily basis and updates once a month the list of MFIs subjected to minimum reserves. However, when comparing this database with the financial information database, there still appeared banks that do exist (they have a website, a location in Google, or have even published an annual report), but they are not registered in the EBA dataset. A clear example is the KNAB bank in The Netherlands that dis missing in the EBA list. It seems that not all banks are included in the EBA list. A limitation of using this frequent EBA-updates is that the banks ceased before 2018 are no longer represented in this dataset, but these banks are still part of the research scope.

4.3. Financial database

The financial database contains information on Revenues, Equity, Total assets, Market capital, Net income, RWA , Branches, Employees, Loans of customers and the Total deposits of customers. All these measures, except branches and employees are measured in million euro's. The retrieved data covers 2016 and 2017. At the time of collecting the data, the latest financial data originated from 2017. When the information from 2017 was not available, data from 2016 is used to fill in. It has been noticed that there is a minimum difference between data from 2016 and 2017, and since it is quantified in millions these small change will not affect the outcome.

The number of online customers could not been found in the external financial information sources. However, this could be an important variable since one of the theories is that adversaries try to victimise as many people as possible to gain their financial rewards. For this reason, the number of customers is multiplied by the percentage of online banking usage from that particular country. This data is retrieved from Eurostat (2018). For example, in Belgium the number of citizens using internet for their banking activities is 67 percent. When a bank is located in Belgium, the Number of customers is multiplied by 67 percent.

4.3.1. Data collection method

There is no open-source or non-commercial dataset that provides financial information for each single European bank. Most of the data provide consolidated data on a country-level. Even the statistics published by the nations Central Banks are consolidated. Often if a nation provides a non-consolidated dataset, the information is limited to two or three bank-size indicators.

The Banker Database is a professional organisation with audited financial reports. The organisation used to provide a trial version for three hours to download their dataset. Unfortunately, this version is not available anymore. A copy of an earlier dataset was made accessible by a researcher of the TU Delft. The following indicators can be found in the dataset: Total assets, Equity, (Net income) and occasionally the Total customers and Total employers. This dataset does not provide all banks nor all indicators required for a solid basis for the financial information dataset.

Financial information will be retrieved from annual financial reports by searching manually through the reports for indicators. Central Banks in countries like Italy, Germany and Spain provide reports for all their individual banks. However, these datasets only provide information on some of the bank-size measures.

4.3.2. Result

The collected data contains information on 2847 individual banks in Europe.

Table 4.1: Coverage of bank-size indicators represented in the financial dataset.

Bank-size indicator	Coverage(%)	Number of values
Revenues	24	683
Equity	46	1303
Total assets	96	2753
Market capital	20	572
Risk-weighted assets	1	33
Net income	27	782
Number of customers	6	171
Employees	48	1362
Branches	45	1292
Loans Customers	55	1584
Deposit Customers	55	1581

4.3.3. Limitation

Some banks provide limited information on the bank-size indicators. Furthermore, the retrieved datasets were mostly in the national language and needed translation. It is hard to say if translated definitions exactly echo the accepted European definitions or if details are ”lost in translation”. The required translations resulted in labour intensive manual work, which may have resulted in an increase in errors.

Furthermore, the limitation concerning the number of online customers: international banks are expected to have a higher percentage online banking users since they can share resources and knowledge. If those international banks are located in a country with a lower percentage of online banking usage, this technique will under-predict the number of online customers. This technique is used as an estimator and is not meant to predict actual numbers.

Reflection on these indicators

The search for data on the selected eleven bank-size indicators caused some challenges. Some of the indicators were named differently and some had different synonyms. Revenues, Net income and Market capital was named differently in many of the financial reports. For example, revenue could be named as turnover, profit, earning and operating incomes. Operating income, however, does not always present revenues very well, but it could still provide a proportional indicator for banks. In the regression analysis this limitation needs to be taken into account.

The theory showed that revenue could be a robust measure of bank-size. However, next to the problem of the inconsistent use of terms, revenue could only be found in a few financial reports and in two country datasets (Italy and Bulgary). It might be the best indicator to represent bank-size from a theoretical perspective, but from a practical standpoint it is hardly possible to find consistent data.

The problem of inconsistent terminology is dealt with by analysing reports to find out what a term exactly means. The manual work has the advantage that you can interpret inconsistent terminology, but the other side of the coin is that human errors can easily be introduced and that it is extremely time-consuming.

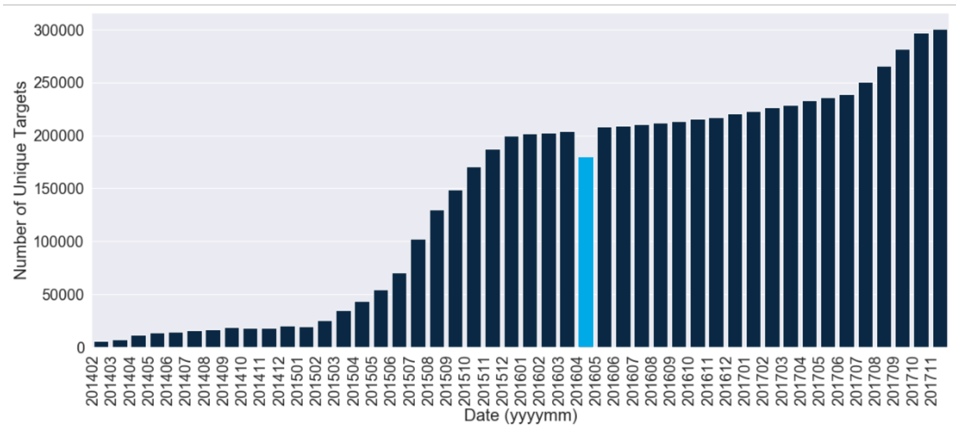


Figure 4.2: Target Frequency over the year 2016 and 2017

4.4. Target database

Fox-IT provides the target database that contains information on the target frequency of European banks. In the Netherlands it is the number one cyber security company that prevents, solves and mitigates cyber threats for government, defence, law enforcement, critical infrastructure, banking and multinational corporations worldwide. The dataset contains information on malware attacks from 21st February 2014 till 3rd November 2017.

4.4.1. Data collection method

The configuration files of the Banking Trojans contain an encrypted list of domains. These domains basically are the targeted financial institutions. These configuration files are analysed and decrypted by a Fox-IT analyst. The targeted domains and other contextual information are saved in the database, called Malware Lab. Since the analysts investigate all types of malware, the database is not limited to certain malware families or clients.

4.4.2. Limitation

Fox-IT analyses and stores the configuration files in the target database. Nevertheless, there is no guarantee that all available configuration file have been found and stored. Moreover, the bias towards client-data cannot be ruled out since the targeted URL has to be linked with the sub(entities). The (sub)entity information is complete for Fox-IT clients, but not for non-Fox-IT clients. The effect on this limitation is reduced in the research of Natalius (2018) through applying domain extraction and matching mechanisms. In this way, more non-customer URls are linked with corresponding bank.

Another limitation is the stability. When observing target frequency (unique count) for all EU countries (see Figure 4.2), it shows a small gap in April 2016 (coloured light-blue). This gap will probably relate to the Dridex group who were very active at that moment. They constantly changed there approach and it took some time to track them when they changed their approach. This causes a little gap of activity in the data.

4.5. Bank characteristics database

The bank characteristics database is retrieved from the research of Natalius (2018), which enables to continue the work of previous scholars in order to develop a comprehensive model. His research has a similar geographical focus and a similar target database. Data on two-factor authentication, the language-use on the banks' website, and domain popularity are retrieved from his database. Below, an overview of how the data is collected:

- **Two-factor authentication**
  The observations for two-factor authentication are limited to the method of *entity authentication*. Entity authentication is concerned with proving the identity of an online banking user, similar to authentication for other online services, such as an email or message(Kiljan, Vranken, & van Eekelen, 2018). The observations are based on the in-log structure, the Frequently Asked Questions page, or documents of the online banking service (Natalius, 2018).

- **Language-use of Banks' website**
  Information about the language-use of the banks' websites is manually collected by browsing to the website and explore which language can be chosen.

- **Domain-popularity**
  Data on domain popularity is retrieved for a list from Cisco Umbrella which ranks the 1 million most popular domains.

- Brand Value

Next to the characteristics retrieved from the research of Natalius (2018), also The brand value ranking data is an extracted list from (Brandirectory, n.d.). Notice that banks from the USA and China are added to this rank but these banks are out of scope and, thus, removed.

4.5.1. Limitation

There are lots of different types of two-factor authentication and some methods will be more easily to detect from the log-in page or from the other available sources. This methods has as limitation that two-factor authentication is in place, but not visible for not-customers. This leads to biases towards a certain type of two-factor authentication or inaccurate data. In addition, the brand value source is not the most reliable data since it is obtained from a website. However, it provides sufficient information on the ranking based on brand value and is easy to obtain.

Summary - Chapter 4

This chapter has set the scope of the research and the data-collection. It summarises the activities done to collect data from several databases and online open sources on the eleven bank-size measures (Chapter 3). Several challenges that to be encountered in collecting data from open-sources to obtain reliable data. Definitions of the measurements differ per bank, making it difficult to obtain adequate and consistent measures. For this reason, conclusions should be drawn carefully regards to **Revenues**, **Market capital** and **Net income**. These indicators are found not to be very reliable. Revenue might be the best indicator to represent bank-size from a theoretical perspective, but it is hardly possible to find consistent and reliable data.

The next chapter will clean and process the data for the purpose of having reliable data for the regression analysis.

5 Data preparation

This chapter elaborates on the data preparation process of the collected data in Chapter 4 required to create a proper database for the regression analysis in the next chapter. To maintain a clear structure, the following four steps of the data preparation are followed: data cleaning, data transformation, data integration, and data replacement. The steps are visualised in Figure 5.1. The steps of data preparation process are executed using the programming language: Python. The utilised and installed packages with their versions are displayed in Table C.1 in Appendix C.

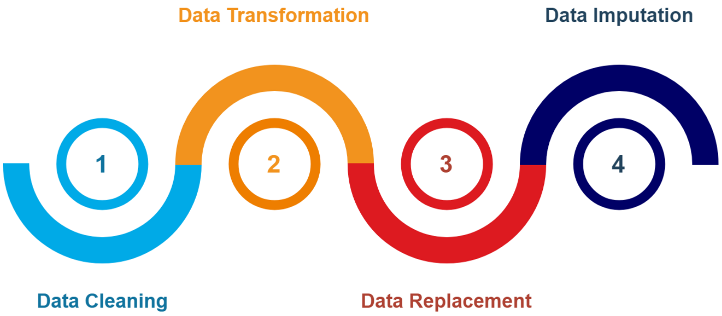


Figure 5.1: The four steps of the data preparation procedure.

5.1. Data Cleaning

This section describes the data cleaning processes for the financial and target database.

Financial Database

- **Checking Bank Existence.** In the past few years, the banking industry has changed, certainly after the economic crisis in 2008. The activities of multiple banks were ceased or taken over by other banks, the government or third parties. To assure that banks who no longer exist will be excluded from the new database, the banks in the financial dataset were compared with the ECB list of banks. If a specific bank has no match with this ECB list, additional analysis needs to conducted. It appears that some banks, i.e. KNAB, already exist for some time but they are not listed in the ECB database. If additional analysis shows that the bank contains information about its location or has delivered a financial report, the bank will be considered and will not be removed from the database. Banks without this information will be removed from the dataset.

  - **Austria - Raiffeissen Group**. The ECB database lists for the Austrian Raiffeisen Group 415 branches. Those branches only have one online banking service and thus have the same domain name. All Raiffeissen branches will be merged and represented by one entity: the Raiffeisen Group.

  - **Portugal-Crédito Agrícola Group**: This group comprises 83 banks. All these banks operate in multiple countries as one universal bank. To ensure consistency in the dataset, these banks will be merged into one bank.

  - The following banks in Spain: ”Banesto, Caja Circulo, Caja de Extremadura, CCM will be merged with other financial institutions. Since there is no data available on those institutions, they will be deleted from the dataset.

- **Check if banks provide online services**. Banks without online services cannot be targeted and will be deleted from the dataset. Natalius (2018) manually checked for every single banks'website in Europe their websites regarding its language-use and two-factor authentication. This dataset will also be used as a guidance to see if the banks provide online services. It appears that some banks not being listed in the Natalius (2018) have now online services. This cleaning procedure will be performed as follows. The banks in the financial dataset and Natalius (2018) dataset are matched. And when no match resembles the respective bank will be manually checked for online services. An additional column in the financial database will then added: 'is_online', where 0 refers to no online services and thus to be removed from the database.

- Banks that were merged or taken over by other banks will be removed from the database. They are considered to be part of the larger group. A few examples are Girocredit (absorbed by Erste Group), Valaris Bank (Wiener Bank), and Artesia Banking (Dexia).

- The available data is also examined on its distribution, outliers and anomalies using a boxplot. A boxplot shows the mean, the median, the Interquartile Range (IQR) and the standard errors. As a result of the manual collection approach for the database, the boxplot shows a few outliers; those errors will be adjusted.

Target database

- Only the year 2016 and 2017 are selected. Argumentation for this selection relates to the financial dataset. The size and target frequency of the bank should be relate to the same time to have valid results. The financial dataset has only data on 2017, whereas data of 2016 is used to fill in the gaps. When looking at the available data, there were only small changes between bank-size measures of 2016 and 2017 that will not effect the outcome. The assumption is that the size of the banks did not change between 2016 and 2017. For the other years, the assumption could not be checked and, therefore ,only the year 2016 and 2017 are selected. This selection ensures that the size matches the frequency of the analysed banks.

- The target database contains information on financial institutions, but also on so called third parties, i.e. SWIFT, Fiducia GAD & IT, PKI or Isabel. Third party companies could be software providers for payment traffic, providers for secure financial messages or online bill payment providers. These parties are not part of the research scope and will be removed.

- Banks outside Europe or banks residing their headquarters outside the EU are not part of the scope and will be removed.

- Branches of cooperations with no information available on an individual level or banks taken over by a group will be removed from the financial database.
- Multiple banks were referred to with a slightly different title but they seem to be the same banks, i.e. Belfius and Belfius bank. In total, 49 bank names are converted into one bank name.

5.2. Data transformation

This section transforms the dataset in such a way that it represents the reality by aggregating the data, it normalised the values, and creates a binary number for being targeted based on target frequency.

Normalisation

Before any selection method or regression can be performed the data needs to be normalised; especially, since the imputation method (KNN-nearest neighbours) will be used which is based on distance calculation. Normalisation prior to the use of this imputation is required. Besides that, the bank-size measures differ in units, ranges and have many variants. To avoid that some features have more influence than others, the data is normalised in a range from zero to one.

Preparation: Binarisation of whether a bank is targeted

The regression analysis investigates which characteristics can explain the target frequency of a bank. The analysis focuses on two possibilities: 1) whether a bank is targeted and 2) how frequently. The first variable is a binary variable of the second. Therefore, a binary 'is_target' variable is created indicating if a bank has been targeted or not.

5.3. Data replacement

A lot of data is missing. The Table 5.1 shows the number of missing values for all bank-size measures. Next to the tabular representation, the matrix (Figure 5.2) provides a graphical insight into the data coverage of the dataset. The black lines indicate the data coverage, while the white spaces show the missing values. The matrix shows that the variable RWA, Number of customers, and Market capital contain limited data. Also, there is a large part where information is available on total assets, loans of customers and deposits of customers, but empty for the other bank-size measurements. This is a result of the German datasets, accounting for 430 banks, containing information on those three measurements. From this figure it is very clear that Total Assets has the highest data coverage in the dataset.

Bank-size Measure	Missing values
Revenues	1556
Equity	929
Total assets	76
Market capital	1624
RWA	1950
Net income	1241
Number of customers	1820
Branches	1013
Loans Of customers	1082
Deposits of customers	1087
Number of Online customer	1804

Table 5.1: The number of missing bank-size values.

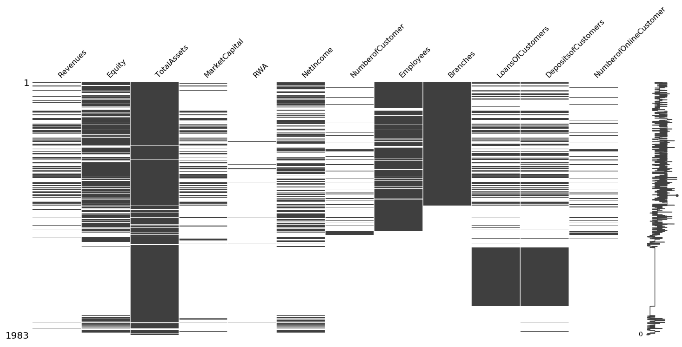


Figure 5.2: Data Density displaying matrix; grey shows when a variable is filled and white reports missing data.

This section clarifies how missing data will be dealt with. Firstly, the missing patterns should be examined in order to understand which interpolation and imputation techniques satisfy the dataset. Secondly, an interpolation technique will be used to predict some of the missing values. Finally, a model-based imputation fills the remaining missing values.

The availability of the banking data differs per country. Most of the countries choose to provide a consolidated dataset on the website of the central bank. Only a few share the unconsolidated financial information about their financial system. The missing observation is, thus, not related to the missing values itself but related to the data sharing values and abilities of a particular country. The relationship between the propensity of missing observation and the values is therefore called **’missing at random’**.

5.3.1. Hybrid approach

The financial dataset contains partial information on the financial measures of almost two-thousands banks. Only six banks have information on all bank-size measures. Using imputation techniques based on six banks to predict all other banks will result in a unilateral dataset. For this reason, an interpolate technique will be applied to predict some of the missing values by linear regression. Linear interpolation is not always the best approach, but the financial dataset is highly correlated and a regression prediction to perform deterministic imputation is a suitable approach to cope with missing data. After the linear interpolation, a multivariate technique will be applied. The idea behind this method is to utilise a reliable approach, namely linear interpolation to fill in the short gaps, which will improve the performance of the multivariate methods.

This approach is endorsed by the research of Junninen, Niska, Tuppurainen, Ruuskanen, and Kolehmainen (2004) who evaluated the performance of imputation methods of missing values in air quality datasets. The missing values of this dataset are defined as missing at random, which is similar to the financial dataset. The research supports the use of a hybrid approach for data imputation. In a hybrid procedure, short gaps are filled by the linear interpolation and the rest of the gaps by a multivariate method.

At the end of the section, the performance of the hybrid approach will be validated by assessing whether the patterns of the generated dataset are similar to the not-processed dataset.

Univariate technique: Linear interpolation

Section 5.3 presented a matrix that helps to identify which indicators were missing but it is also useful to recognise which indicators are not overlapping, which could help predicting missing variables of other indicators.

Interpolation is an approach to cope with missing data. It calculates an algorithm on known data and replaces the missing data with predicted values based on the calculated algorithm. Different techniques can be used to interpolate data, i.e. linear, cosine, cubic or hermite interpolation. Since the bank-size measures are highly correlated it is expected that linear regression will be the best fit to predict the missing values. The Python interpolate package will be used to perform the linear interpolation. Other potential methods are: quadratic, cubic, akima and pchip. However they did not give the desired output.

When utilising linear interpolation one variable needs to be able to predict the other missing values. Three guideline rules are suggested when choosing the predictive variable: 1) strong correlation - correlation coefficient > 0.8, 2) significance - p-value < 0.05 and 3) ability to predict correlation.

The *Pearson’s correlation coefficient* is a statistical measure for the strength of a linear relationship between paired data. The magnitude of the correlation coefficient determines the strength of the correlation. Scientific studies argue what defines a strong correlation. This re-

Table 5.2: Rationale of imputed bank-size measurements

Imputation variable (Bank-size measure)	Predictive variable	Correlation	P-Value	Ability to Predict
Revenues	Total assets	0.8568	0.0015	1489
Equity	Total assets	0.7194	0.0261	862
Net income	Revenues	0.8678	0.0282	96
Branches	Deposits of customers	0.8150	0.0480	491
Deposits of customers	Branches	0.8150	0.0480	565

Table 5.3: Rationale of non-imputed bank-size measurements

Bank-size measure	Reason not able to be predicted
Total Assets	There are no variables that have data on the missing values of Total Assets.
Market capital	Only RWA , correlates relatively high (0.75) with market capital, but is not able to predict. Only revenues had a moderate correlation level, others were week or very weak. However, also revenue did not have data-points that market capital had not.
RWA	RWA will not be analysed (to less data points and is part of total assets) predicting using total assets will generate biased results.
Number of customer	All have high p-values, expect for deposits of branches, but had a weak correlation value.
Employees	All p-values are above 0.58, thus not significant.
Loans of customers	Does not have a predictor that fulfils all three criteria.
Number of Online customers	Similar to number of customers.

search follows the guideline of Evans (1996). He suggests the following for the absolute value of r; 0.00-0.19 'very weak', 0.20-0.39 'weak', 0.4-0.59 'moderate', 0.60-0.79 'strong', 0.8-1.0 'very strong'. The measures that correlate above a correlation coefficient of 0.8 is evaluated by Evans (1996) as very strong. Even though scientific studies argue on what defines strong correlation, a correlation of 0.8 is enough to state that there is a (very) strong correlation and will therefore be chosen to be the threshold value. The relation should be *significant*, where the standard significance rule of a p-value lower than 0.05 is demanded. The *ability to predict* means that the predictor needs to have information on what is missing. For the bank-size measure the value will be predicted (interpolation variable). This information can be found in Figure 5.2. The predictor should contain information (grey coloured) which is missing for the potential interpolated bank-size measures. An overview of the selection criteria for each bank-size measure that fulfil all three criteria is shown in Table 5.2. Table 5.3 provides an overview of the reasons why some bank-size measures could not be interpolated. The visual graphs of the data points and the generated interpolation points can be found in Appendix D.

Limitation

The limitation of the linear interpolation is that the predictive missing values cause over-identified relationships since the fitted data does not have an error term included in their estimations. The regression model predicts the most likely value of missing data but does not provide uncertainty about that value. Consideration whether limitation influences the nature of the data set will be explored in the next section.

5.3.2. Data imputation

Given that the financial dataset still misses many values after the imputation process, the dataset is completed by using multivariate nearest neighbour (NN) method by Dixon (1979). The algorithm searches through the dataset and looks for the most similar instances. The similarity of two instances is determined using a distance function (Acurna & Rodriguez, 2004). For this research, it is chosen to apply the Euclidean distance.

The missing value are replaced be using the mean value of the attribute in the k-nearest neighbourhood. The choice of the number of neighbours is K. It is difficult to determine the most optimal K. A small K will result in overestimating a few dominant instances and a large K could be counterproductive because the noisy data will also be taken into account (Acurna & Rodriguez, 2004). Most researchers argue that the higher the number of neighbours, the fewer errors, where the number 10 seems to perform well for many problems (Kohavi et al., 1995). In this research, the elbow method, a python implementation, is used to define how many neighbours facilitates the optimal fit concerning the dataset. A visualisation of the elbow method can be found in Figure 5.3. The optimal value of k is were the improvement of errors declines, thus were the graph creates the elbow shape. In this case, K = 4 shows the most optimal option, and will therefor be used in the k-fold cross validation.

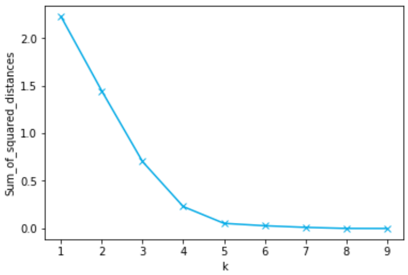


Figure 5.3: Visualisation of Elbow Method for most optimal K

limitation
Aforementioned, the main limitation of k-fold cross validation is choosing the best value of K to avoid overestimating dominant instances or overfit the data as an effect of the noise in the data. Since the financial dataset is manually collected, there is noise in the dataset, and, thus, it is likely that the data is overfitted.

5.4. Data merging

The final step of the data processing procedure is to combine all collected datasets into one. Figure 5.4 shows how the three tables: financial database, target database and the other target selection characteristic tables are integrated. From this visualisation it is clear that the tables are linked based on the Bank Name and ID. However, the names and order of the names of the banks slightly differ in the databases. Additionally, the bank ID's were created in the data sheet itself and did not match. These dissimilarities made it a challenge to link the tables. Consequently, a linking table is created to link the Bank Names in the target database with the financial database and the other target selection characteristics. The Python FuzzyWuzzy package is used to find, for each listed bank in the target database, the corresponding bank in the financial database. FuzzyWuzzy is a string matching package. The calculations show how well the string matches through calculating the Levenshtein distance. The matches identified by FuzzyWuzzy are manually checked to ensure that all banks have been matched correctly. Banks using abbreviations in their names could not be processed with the python package and needed manual checks and corrections.

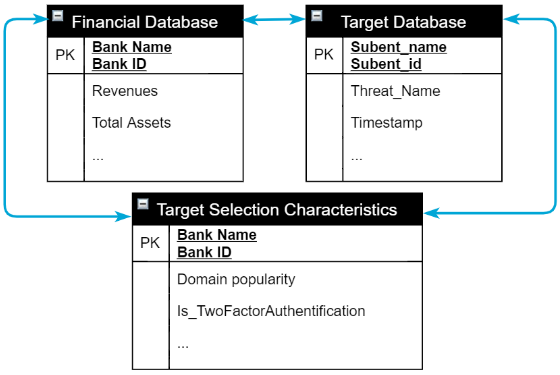


Figure 5.4: Visualisation of database merging

Summary - Chapter 5

Previous chapter collected data of 11 bank-size measures for 1400 banks from open sources. The data was not fit for purpose yet and had to be cleaned in order to make the set reliable and valid for further analysis. This is done by the following four steps of a iterative data preparation procedure: data cleaning, data transformation, data integration, and data replacement.

The dataset is characterised with a high correlation, the missing values are at random and only six rows contain full information. Given these properties, an attractive imputation method to use is a hybrid approach, combining linear interpolation with multi-univariate imputation. The linear interpolate was used to fill the short gaps in the database. The multivariate K-nearest neighbours imputation technique was used to complete the dataset.

The next chapter provides a summary of the data through a descriptive analysis. In addition. it aims to seek insight into the relationship of the various banks' characteristics with each other and with (the number) of being targeted.

6 Descriptive analysis

After cleaning the data in the previous chapter, the data is ready for the analysis. This chapter provides a summary of the data sample. It aims to seek insight into the relation of particular banks' characteristics by creating visualisations. All visualisations are presented in Appendix E, however, the most important outputs are described in this section.

Before the descriptive analysis can be conducted, it is vital to determine how to measure target frequency. Various metrics can be used to measure this. The metrics are examined on their advantages and disadvantages in order to determine which one fits the purpose of this research best.

6.1. Measuring target frequency

In literature, there are three metrics identified that can measure the target frequency: Raw count, Week count and Unique count. These metrics are briefly described below.

- **Raw count** counts how many times a specific bank domain name is listed in the configuration file of the malware.

  Advantage: it is an easily observable and computational metric and is also a widely used metric for security reports.

  Disadvantage: according to Tajalizadehkhoob et al. (2014) and Natalius (2018) this might not be a good metric of measuring the relative degree in which a domain is targeted since it is highly dependant on how many files are distributed. The configuration files are, depending on the adversaries intent, updated after a certain period. These updates could occur after eight months or every two hours. The Raw count metrics counts each of these configuration files as a different target which is in reality not always the case. An example is TheTrick which is an active malware that only changes the IP-address in the configuration file of the fake website where the victim is redirected to. This metric recognises this as a new target, but in reality this is not a new target..

- **Week count** counts the number of weeks a bank is listed in the configuration list of the malware. It sums all targets in a week-interval. Week count assumes that in a week-interval the adversary will update the list followed by new targeting.

  Advantage: This metric is proposed by Tajalizadehkhoob et al. (2014) to avoid overestimation of Raw counts.

  Disadvantage: The configuration file is updated based on the adversaries decision. Counting based on a week-interval does not well represent the number of targets; if an

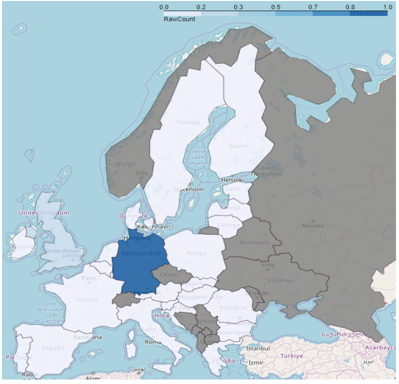


Figure 6.1: Geographical distribution of targets (Raw count)

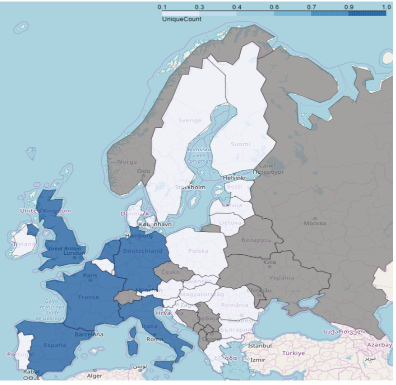


Figure 6.2: Geographical distribution of targets (Week count)

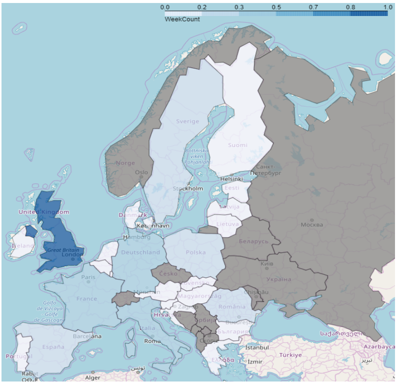


Figure 6.3: Geographical distribution of targets (Unique count)

adversary updates its configuration file after two weeks, it doesn't automatically mean that a new target has been identified. It could easily be that a domain or a redirected IP is added/removed/changed. A week-interval does not take this into account. Natalius (2018) confirms this and mentions that Week counts better deal with overestimation, but that still many conditions need to be checked before applying.

- **Unique count**. The Fox-IT target table generates a Unique Attack ID considering the following unique variables: target (domain), inject code, attack type, time and configuration file. When an adversary uses different codes and methods, it is assumed that this attempt could count as a new target. In this way, when looking at the inject code instead of counting the configuration files, Unique Count is able to differentiate a new target from an update.

  Advantage: According to the research of Natalius (2018) and assuming that different inject codes indicate different attack attempts, the metric can significantly eliminate the overcounting problem of the Raw count metric. Besides this, it is not based on the unknowable assumption that an adversary is weekly updating its configuration files. The count from this metric denotes the actual number of targets on a bank.

  Disadvantage: Different malware families use similar types of attack, i.e. fake web inject. Different malware families are thus categorised under the same Unique Attack ID since their inject code is similar. Adversaries behind the different malware types differ, and they are in charge of which domain is listed in the configuration. It is not valid to assume that when adversaries target the same domain and use the same attack type, this can be counted as one attack.

Evaluating the target frequency metrics

Figure 6.1, 6.2 and 6.3 show the country distribution of targets utilising Raw Count, Week Count and Unique Count. The maps show that each metric leads to a different geographical distribution. Similar to bank-size measures, the target selection metric influence the outcome and thus the analysis. The metrics are evaluated in order to select the best measure for the analysis.

When counting the target frequency in order to differentiate it is important to know what the adversary is doing: targeting or updating files i.e. redirecting towards a new IP-address etc. The Raw count does not make the distinction between targeting and updating and will therefore not be used in this research. Although Week count makes the distinction between weekly targeting and updating, it has been noticed in the target database that updating configuration files also happens after months and that this is not limited to a week-interval. The Unique Count is able to make the distinction between updating and targeting by looking at the inject-code. Although the Unique Count doesn't take the different malware types into account, it is possible in Python to count the Unique Attack ID and Threat, presenting a valid metric. This updated version of Unique Count is applied in this research and will be further referred to as Unique Count.

6.2. Evaluating the focus of banking malware

This research also aims to understand the underlying reasons for the narrow focus of banking malware. Tajalizadehkhoob et al. (2014) and Natalius (2018) show that a small percentage of banks are subject to more than 80 per cent of the targets regardless of the metrics that are used. Below, the dataset is examined, using various metrics to confirm if this narrow focus is also visible in the dataset used in this research.

The Empirical Cumulative Distribution Function (ECDF), figure 6.4, provides a broad picture of the dataset. The plot represents the *normalised values* of target frequency on the x-axis and the percentage of banks on the y-axis. The plots distinguish the three metrics that count the number of targets: Raw count, Week count and Unique count.

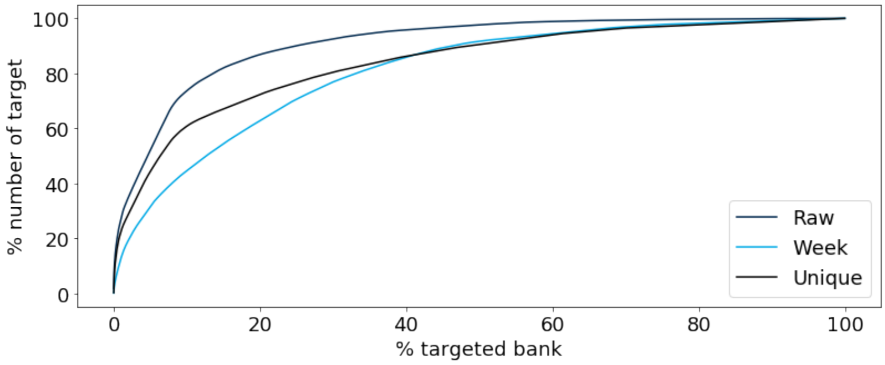


Figure 6.4: Cumulative distribution of targeted banks

The ECDF (Figure 6.4) proves that a small percentage of banks count on to a significant proportion of targets. To what extent the power law distribution is applicable differs per approach of counting. Using Raw count, 20 percent of the banks attract more than 80 percent of all targets, confirming the statement of previous research. However, when using Week count, 20 percent of the targeted banks account for 60 percent of all targets. Moreover, utilising Unique count, 20 percent of all attacked banks account for 70 percent of the targets.

In this dataset, covering target selection of the year 2016/2017, the adversaries focus towards certain banks is not as high as previous studies mention. However, when using Unique count, still a small percentage of banks (20%) attract a significant number of targets (70%).

6.3. General description of the data

The merging of the datasets in the previous chapter results in information about 1293 European banks, 887 targeted banks and 406 non-targeted banks for the year 2016 and 2017. The dataset contains information about the following bank characteristics: eleven bank-size measures, domain-popularity, brand-value, language-use on websites, being part of a banking group, and two-factor authentication.

Below, visualisations show the number of banks per country, the target intensity towards countries and the malware threats in order to provide a deeper insight into the data.

Banks per Country

As mentioned during the data collection, some countries have many small banks, whereas others have a few large banks. Figure 6.5 roughly shows the number of banks for each country.

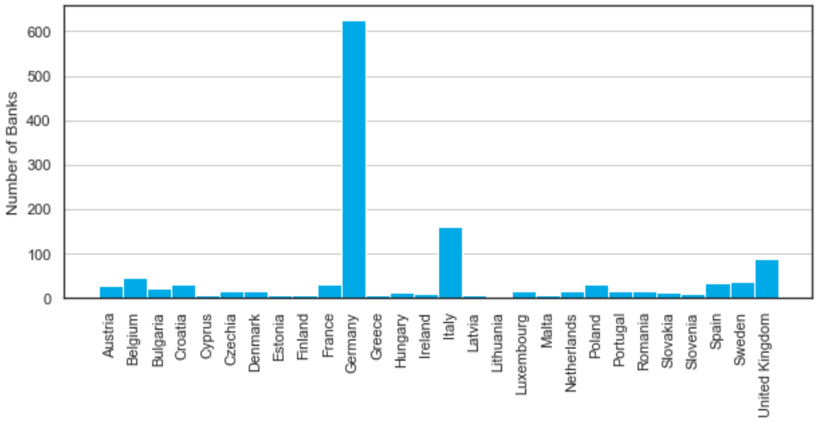


Figure 6.5: The number of banks per country

The figure shows that Italy has more than 100 banks and Germany more than 600 banks. The high number in Germany is caused by its many savings banks, the Sparkasse, which already account for 431 banks.

Intensity of targets per country

The total number of targets in 2016 was 10913, and in 2017 it was 10827. Figure 6.6 shows the normalised number of unique targets over the total number for the year 2016 and 2017. The figure shows that Germany, France and Italy have the largest number of targets in 2016, but these numbers decrease in 2017. In some other countries, this movement is visible in the opposite direction and increased in 2017: Denmark, Finland, France, Luxembourg, Spain and Sweden. The expected trend towards targeting East European banks is not visible in this data.

Figure 6.7 shows the number of attacks normalised over the number of banks. Here, we see that German banks show a low target intensity per bank, whereas Finland and Ireland have a high concentration per bank. Although not all banks are included in this data, in proportion they represent the reality (Chapter 5.2), and provide understanding which countries are subject to the most targets.

Malware threats

A description of the malware threats in this database can be found in Chapter 2 and Appendix A. The malware types available in the dataset are: 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.

Figure 6.8 shows the activity of different malware types over time. Some are active for a long or short time, and some are reactivated later to be used again. Some observations from this figure are that Tinba v1, GootkitLoader and Corebot are identified only once in 2016. Bokbot is a new malware discovered in November 2017. TheTrick, Dridex-Loader,

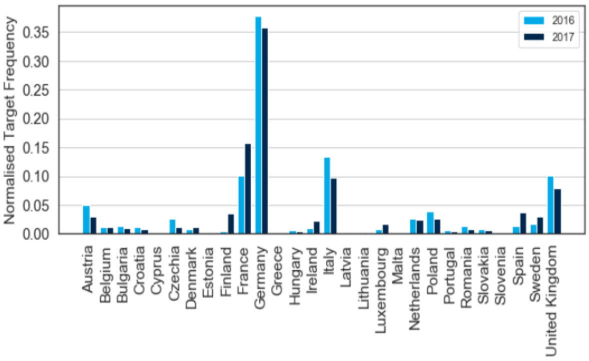


Figure 6.6: Number of normalised targets per country

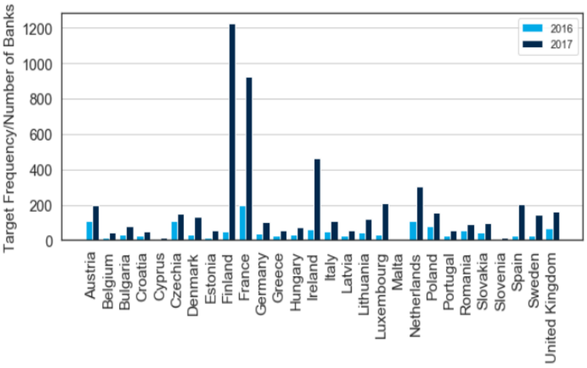


Figure 6.7: Number of normalised targets per bank per country

Zeus-Panda, KINS, Gozi-ISFB and Citadel are longterm active malware threats.

Figure 6.9 shows the number of targets from certain malware threats in a specific country. The values are normalised over the total number of targets per malware threat in order to identify in which countries the malware threats are most active. The figure tells us that different malware types frequently target Germany. Also is visible that Nymaim and Pkybot mainly targeted the United Kingdom, Qakbot and ReactorBot the Netherlands, Zeus Action France and Matrix targeted Poland.

Zeus was discovered in 2007, but variants such as Zeus-OpenSSL are still active. Figure 6.10 shows the number of targets per malware threat. Gootkit, Zeus-OpenSSL's and Dridex-loader are active with over 2000 banks.

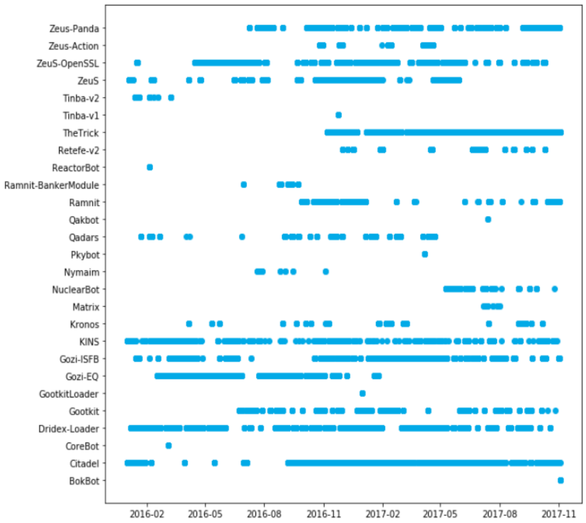


Figure 6.8: Types of malware threat over time.

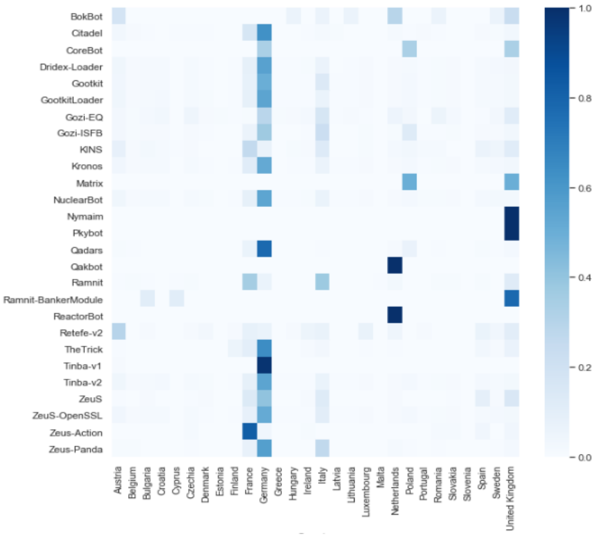


Figure 6.9: Target Frequency per malware threat

6.4. Insight into the characteristics of targeted banks

This research aims to add bank-size to the model, and therefore, bank-size will be described in detail. Also insight into the relation between the other bank characteristics and whether a bank has been targeted are provided.

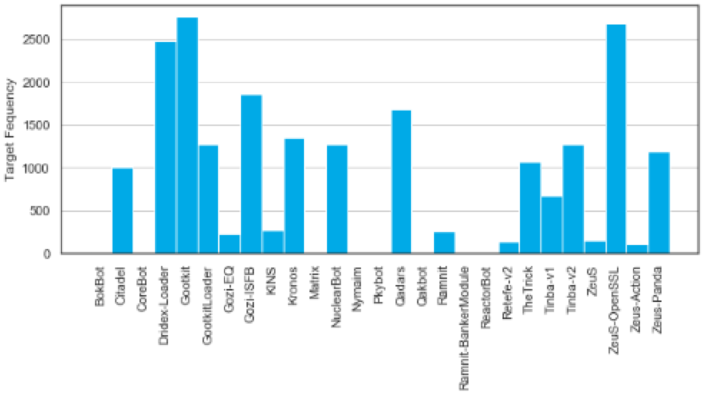


Figure 6.10: Types of malware threat over time

Bank-size

All bank-size data-points in this section are plotted into two categories: targeted and non-targeted in order to gain deeper insight into the size of targeted and non-targeted banks. The graphs for every single bank-size measure can be found in Appendix E. From these plots, it can be concluded that expect for the number of online customers, **non-targeted banks are smaller in size** and **regardless of the bank-size measure, the largest banks have always been targeted**.

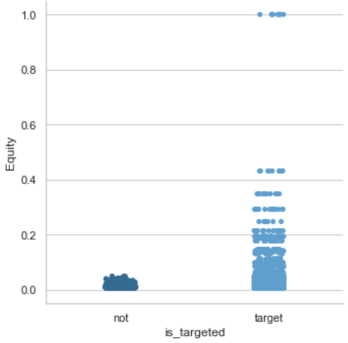


Figure 6.11: Equity of targeted and non-targeted banks

.

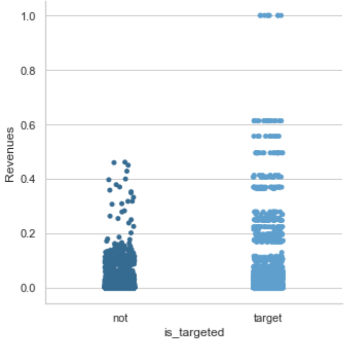


Figure 6.12: Revenue of targeted and non-targeted banks

.

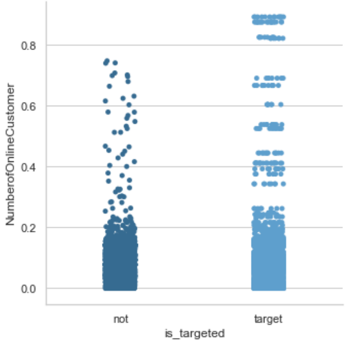


Figure 6.13: Number of online customers of targeted and non-targeted banks

.

This statement is confirmed in figure 6.11. It shows concentrated equity values for non-targeted banks. A clear pattern is visible where all non-targeted banks have a normalised equity-value below 0.2. Not only equity but also total assets (Figure E.4), market capital (Figure E.3), number of employees (Figure E.5) and loans of customers (Figure E.2) show the same pattern where all non-targeted banks have normalised values below 0.2.

Less concentrated values, but still providing a clear pattern where all non-targeted banks have normalised values below 0.5 is shown in Figure 6.12. Other bank-size measures showing the same pattern are: deposits of customers (Figure E.8), net-income (Figure E.7) and number of customers (Figure E.6).

The trend where all small banks have normalised values below 0.5 or even 0.2 is not visible for the number of online customers. This computed measure does not provide a clear pattern; not-targeted banks and targeted banks have both high values. This is visible in Figure 6.13.

Besides indicating that non-targeted banks are the smaller banks, from the plot is also very clear that the largest banks (normalised value of 1.0) are always targeted.

Other banks' characteristics

The mean of every bank-size measure for targeted and not-targeted banks has been plotted to seek insight into the difference of bank-size of targeted versus non-targeted banks. Figure 6.14 confirms the above made statement. Except for the number of online customers, the mean of each bank-size measure is larger for targeted banks. It also shows that targeted banks have a higher average of domain-popularity, being more often part of a banking group, and have a higher brand value. Besides, targeted Bank have on average less two-factor authentication compared to non-targeted banks.

Figure 6.15 and 6.16 show that some languages score on average higher for targeted banks compared to non-targeted banks. Here, the Bulgarian language shows an a-typical result, having almost a negligible value for non-targeted banks.

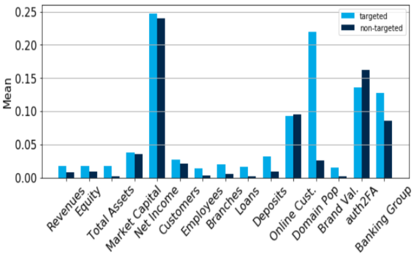


Figure 6.14: Mean of bank-size, Auth2FA, brand, for (non-)targeted banks

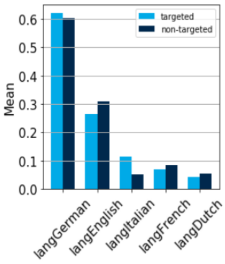


Figure 6.15: Mean top 5 languages for (non-)targeted banks

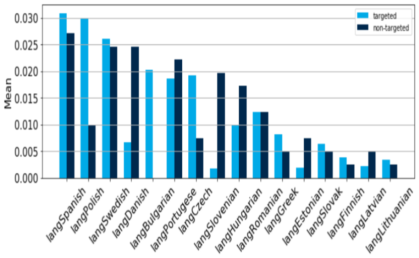


Figure 6.16: Mean other languages for (non-)targeted banks

.

.

.

6.5. Insight into the banks' characteristics related to target frequency

This section seeks to provide insight into the relation of the size of a bank and its target frequency. Firstly, the banks are divided into the target frequency: low (<33%), medium (33-67%) and high (> 67%). Per category the mean of each bank-size measure is visualised in Figure 6.17. From this graph, it is clear that except for the number of online customers, the mean of each bank-size is higher when the banks is more frequently target.

Secondly, the banks are divided based on their median in large banks (light blue) and small banks (dark blue), see Figure 6.18. Herein, the average number of targets for small and large banks are visualised. Noteworthy is that large banks in terms of bank-size (except: employees and the number of online customers), brand value and a high domain popularity are more often targeted. Large banks having high deposits/loans of customers, revenues or equity are twice as many times targeted compared to smaller banks. From Figure 6.19, it is visible that the absence of two-factor authentication leads to more targets and the presence of the Germany, Greek, Estonian, Slovakian, Latvian, and Lithuanian language leads on average to more targets.

6.6. Understanding the relationship between bank' characteristics

The previous section showed that the financial dataset correlate; e.g. when a bank has more customers, it will likely have more deposits. However, also other characteristics influence each-other. For example; when large banks have more money to spend on security personnel and on IT security measures this may implicate that large banks will have more often two-factor authentication compared to the smaller banks. It must be investigated, before running the regression model, how those characteristics relate to each other.

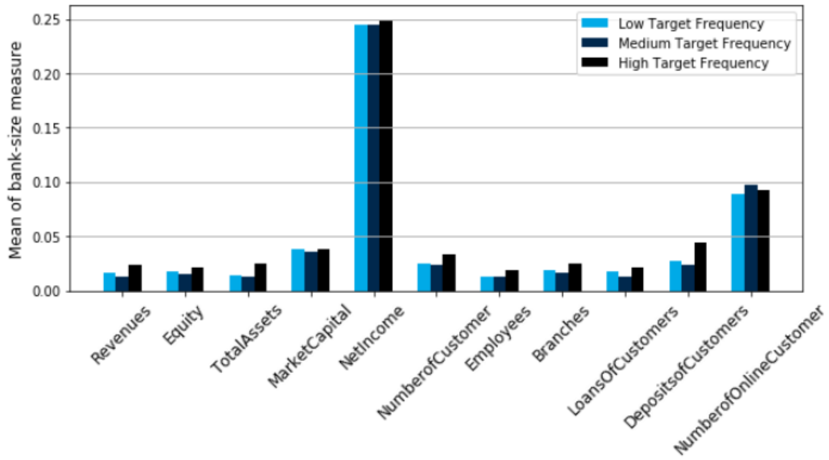
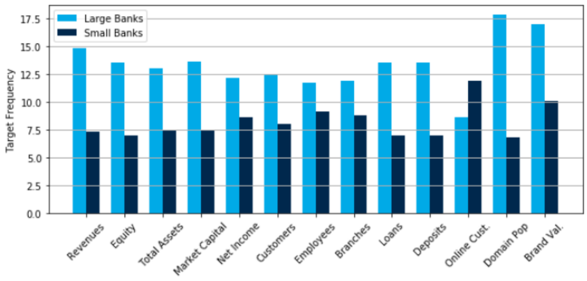


Figure 6.17: Bank-size in relation to low, medium and high target frequency



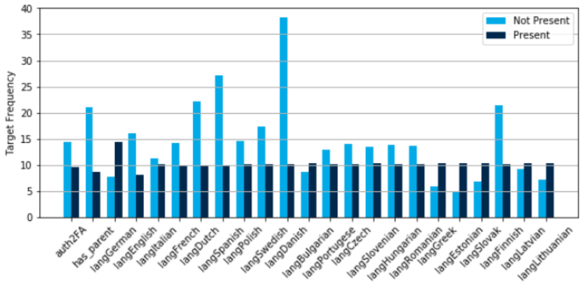


Figure 6.18: Mean of target frequency for continuous variables
.

Figure 6.19: Mean of target frequency for dummy variables
.

Figure 6.20 and 6.21 show a correlation heatmap. The values in the heatmap shows how much the variables correlate. The higher the correlation, the darker the colour. Notice that all negative numbers are coloured light yellow, but this does not indicate a low/large correlation.

Appendix E provides the reader with the complete correlation matrix with all variables. This correlation matrix supports the identification of relationships between the banks' characteristics and being targeted/its frequency. Below, only the outstanding results are discussed supporting a better understanding of the outcomes of the regression analysis.

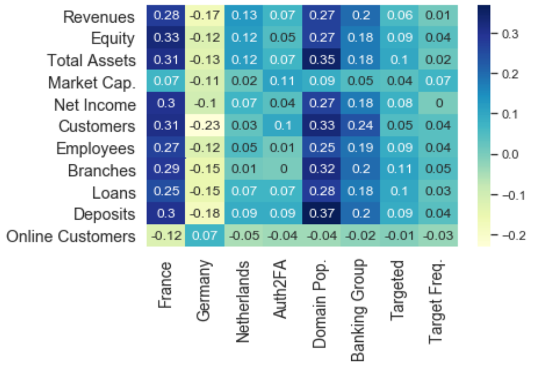
General Findings

- Except for market capital and number of online customers, all bank-size measures are moderate to high correlated. (Figure E.11)

- Bank-size is strongest correlated with domain popularity. An expected result since popular banks are larger (Figure 6.20).

- French banks are the largest, the Dutch banks are the second largest and German banks are the smallest in terms of size. In addition, being part of a banking group correlates the strongest with the French language. Thus, France has relatively large number of internally operating banks. (Figure 6.20).

- The most banks with the two-factor authentication implemented are **high**: Sweden, Belgium, Croatia, Denmark (and Latvian). And extremely **low**: Germany (Figure 6.21).



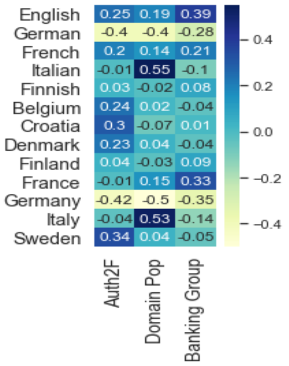


Figure 6.20: Bank-Size related to banks characteristics, country and target selection

.

Figure 6.21: Two-factor authentication, domain pop., and banking group related to country/language

.

- There is a moderate relationship between the Italian language and the domain popularity. It appears that many customers operate online for their banking business (Figure 6.20).
- The French and Dutch language moderately correlate. The reason underlying this correlation is that Belgium has officially three national languages (French/ Dutch/ German), whereas the most used are French and Dutch. This relation is introduced by the Belgium banks (Figure E.11).
- The provision of English language on a website does, apart from the UK, relate the strongest to being part of a banking group. This is in line with the assumption that those banks operate internationally. Also, banks providing the English language are more likely to have two-factor authentication and have a higher domain-popularity score (Figure 6.21).
- For the visibility factors, brand value, domain popularity and being part of a banking group, they correlate negligible with each-other (Figure E.11)
- The German language negatively correlates with domain popularity, two-factor authentication and English. (Figure 6.21).

Bank characteristics related to Being Targeted/Target Frequency (Figure 6.22)

- Being targeted is negligible correlated with the target frequency (0.24).
- Bank-size is (negligible) positive correlated with being targeted and its intensity.
- Domain popularity strongly correlation with being targeted.
- Italy is after domain popularity the strongest predictor.

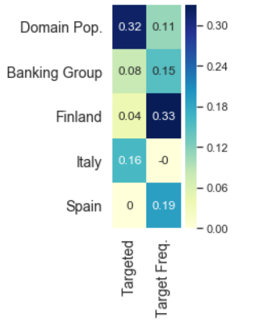


Figure 6.22: Most correlated features in target selection

- Finland and Spain are the strongest predictors for ”being targeted”. Note that the intensity is more related to geographical location, but according to Figure 6.6, we see this changing over time. Thus, it does not mean that Spain and Finland are generally speaking a strong predictor; it only applies for 2016/2017.
- Being part of a banking group is more important predictor than domain popularity for the intensity of targets.

Summary - Chapter 6

This research uses the Unique count to measure target frequency. It is able to make the distinction between the updating of a configuration file and real targeting by using the inject-code. In this dataset, covering the target selection of the years 2016/2017, the adversaries focus on specific banks is not as high as in previous studies mentioned. Still, a small percentage of banks (20%) is subject to a significant number of targets (70%). The descriptive analysis shows that, except for the number of online customers, non-targeted banks are smaller in size and the largest banks have always been targeted. In addition to that, the use of the two-factor authentication still leads to being targeted but lower in frequency. The correlation matrices shows that the banks' characteristics are highly correlated. Two-factor authentication correlates with a variety of countries and specific languages. Bank-size highly correlates with domain-popularity.

The next chapter establishes the regression model and deals with multi-collinearity. The problem of multi-collinearity is expected as a consequence of high correlation in the bank-size measures identified in this chapter.

7 Regression model and results

The regression model aims to reach conclusions beyond the descriptive analysis in previous chapter. In this chapter, two types of regression analysis will be performed: a Logistic Regression and a Negative Binomial Regression. The Logistic Regression examines significant banks' characteristics associated with being targeted. The Negative Binomial Regression explores which banks' characteristics can predict the target frequency. The regression model is formulated and analysis in the programming language R. The utilised and installed packages can be found in Table C.2 in Appendix C.

First, the respective dependent and independent variables of both models are described. Secondly, a Principal Component Analyse (PCA) is preformed to deal with multicollinearity in the regression model. And finally, the results of both regression models are discussed.

7.1. Variables of Regression Model

This section describes the dependent and independent variables for the Logistic and Negative Binomial model.

7.1.1. Dependent variables

The dependent variable for the Logistic Regression is the dichotomous variable: being targeted. For the Negative Binomial Regression the dependent variable is the continuous variable: target frequency.

7.1.2. Independent variables

The independent variables for the Logistic Regression and the Negative Binomial Regression are similar and are summarised below:

- **Bank-size**; these are normalised (between zero and one) continuous variables.
  Variables are: Revenues, Equity, Total assets, Market capital, Net income, Customers, Employees, Branches, Loans of customers, Deposits of customers and Online Customers.
- **Language-use on a banks website**; is a boolean variable related to the presence of a language used on the banks' website.
  Variables are: English, German, French, Dutch, Italian, Spanish, Portuguese, Greek, Czech, Slovak, Slovenian, Polish, Hungarian, Romanian, Bulgarian, Danish, Swedish, Finnish, Latvian, Estonian and Lithuanian.
- **One-factor authentication**; a boolean variable representing if a user needs to identify him/herself with a username and password.

- **Two-factor authentication**; a boolean variable assigning if two-factor authentication is in place. Two-factor authentication is a second security layer to confirm the users' identity next to the username and password. This could be a security token (PIN) or biometric measure.

- **Domain popularity**. The popularity algorithm is the number of unique client IPs visiting this domain, related to the sum of all requested domains. An inverse ranking is applied; the highest rank get highest score. When a domain of a bank is not ranked in Cisco's list, a zero is imputed. The domain popularity score is a continuous variable.

- **Brand value ranking**: is a ranking based on the brand value from the top 500 banks in 2017. The ranking applies the same inverse ranking method as domain popularity.

- **Part of Bank Group**: when a bank is part of a bank group a one is assigned and when this is not the case a zero is assigned.

- **The Control variables** serve as a control mechanism to reduce biases in the regression model. These variables are included to avoid that, for example, a certain language only influences target selection as a effect of the significance of the country - speaking that particular language, to target selection. The following control variables are added to the model:

  - **Country**; the country were the bank is located (Logistic and Negative Binomial).
  - **Language count**; refers to the number of languages offered by a bank on their online banking site (Logistic and Negative Binomial).
  - **Threat**; the malware variant of the target (Negative Binomial).
  - **Year**; the year when the bank was targeted (Negative Binomial).
  - **Unique URL count**; the number of unique URLs corresponding to the online banking entity in the target database.

7.2. The Multicollinearity problem

According to the requirements of a regression model, the independent variables should not relate to each other Ilvento (n.d.). From the descriptive analysis in Chapter 6 it is known that the bank-size measures highly correlate, this is shown in the heatmap in Figure 7.1. When a particular variable, such as Total Assets, is highly correlated with a set of independent variables - the bank size measurements, **multicollinearity** arises. (Ilvento, n.d.). Even though multicollinearity does not affect the predictive power or reliability of the model, it can effect the precision of the coefficient estimates, resulting in unstable and poor estimators (Ringle, Wende, & Becker, 2015).

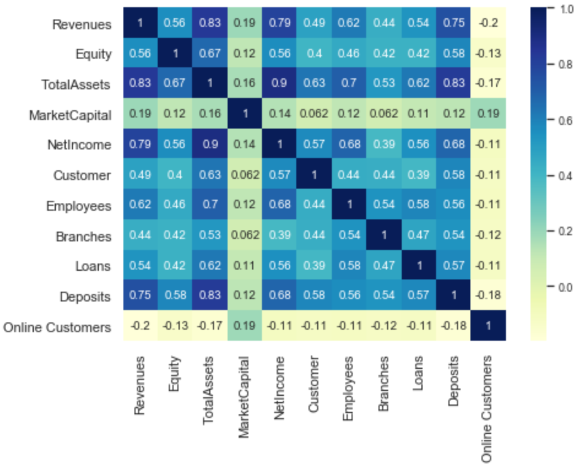


Figure 7.1: Correlation Matrix; bank-size characteristics

7.2.1. Approaches

There are various approaches to reduce multicollinearity. The first option to deal with it in a regression model is to add interaction terms to the model. However, as de González, Cox, et al. (2007) state, the difficulty of interpreting interactions rapidly increases with the number of factors involved. With eleven bank-size measurements, multiple high-order interactions should be added to the model. Since the bank-size measures not only show multicollinearity with each other but also correlates with other independent variables, it is not possible to adequately add interaction terms given the number of variables. Interaction terms will result in an unnecessary complex model.

Secondly, feature selection analysis can be applied to remove irrelevant and redundant information on the bank-size measures. Excluding irrelevant or distracting variables for measuring bank -size results in less collineartiy. Feature selection is performed and can be found in Appendix F. Multicollinearity can be measured using Variance Inflation Factor (VIF). After this approach, the VIF appeared to be smaller and was good enough for the regression analysis.

Last, Principal Component Analyse is able to transform eleven bank-size measures into one ”bank size” variable, making it less complicated- compared to feature selection and interaction terms, and reduces multicollinearity. It is chosen to apply PCA to deal with the problem of correlating independent variables. The analysis will be explained in the next section.

7.2.2. Principal Component Analyse

The goal of this section is to derive one single ”bank size variable” from all eleven bank-size variables. The choice has been made to use PCA since it is a well-known general mathematical tool for dimensional reduction, is implemented in R and Python, the programming language practised in this research and, lastly, it is fast to use.

PCA is capable of transforming many correlated variables into a smaller number of uncorrelated variables called **principal components**. In this way, the analysis extract important information from the dataset and presents it as a new non-redundant or non-overlapping components(Abdi & Williams, 2010). These components are a linear combinations of the original bank size variables weighted by their contribution to explaining the variance of the dataset (Cangelosi & Goriely, 2007). The first component is defined as the linear combination preserving as much as variation as possible among all linear combinations of the bank size variables. The second principal components capture the highest variance from what is left after what is explained by the first component.

Two types of datasets are used to perform the regression analysis; the Logistic Regression dataset and the Negative Binomial Regression database. The difference between those datasets is that the logistic dataset contains information on non-targeted banks and the Negative Binomial Regression doesn’t. Therefore, a PCA is performed for both datasets. The principal components are calculated using the built-in R function to perform the PCA. The ”factoextra package” (Appendix C) is used to visualise the PCA analysis. The outcomes of both PCA analyses can be found in Appendix F.

Aforementioned, the first component explains as much variance as possible. The explained variance for all components for both datasets are visualised in Figure 7.2 and 7.3. These figures show the percentage of variance explained by each principal component. In this figure it is visible that the first components explain around 55 percent of the variance and the second component over 10 percent, together they explain around 65 percent.

To get more insight into the relation between the bank-size variable and the principal components, a graph is created (Figure 7.2 and 7.3). In this graph, the bank-size variables are plotted on the first component (x-axis) and the second component (y-axis). Here, the correlation between a variable and a principal component is used as the coordinates of the variable on the graph (Abdi & Williams, 2010). The plot can be interpreted as follows; the variables on the right side of the graph, every bank-size variable - except for the number of online customers- point to the same direction, these are all positively correlated variables.

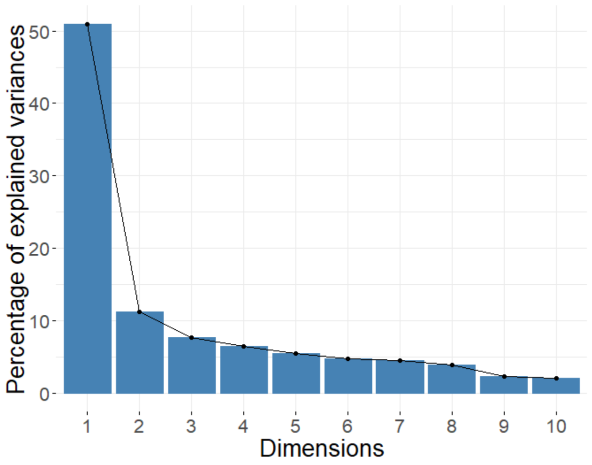


Figure 7.2: Explained variance for each calculated component - Logistic database

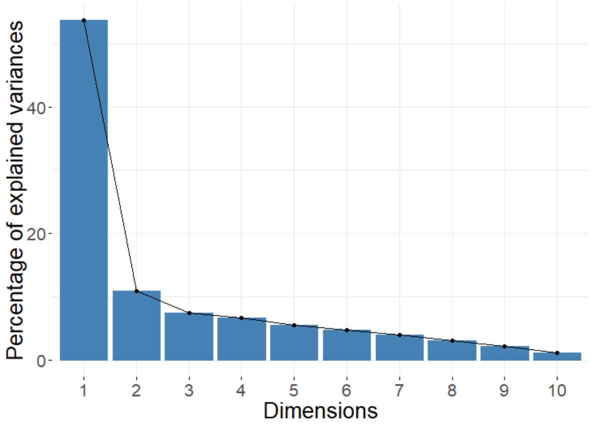


Figure 7.3: Explained variance for each calculated component - NB Database

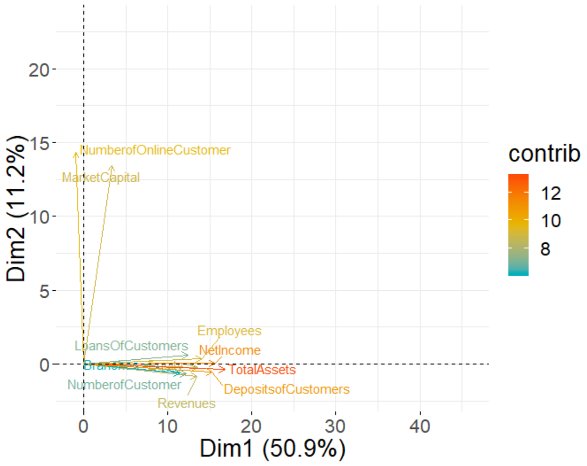


Figure 7.4: Bank-size Variables plotted on the two largest variance explained components - Logistic Database

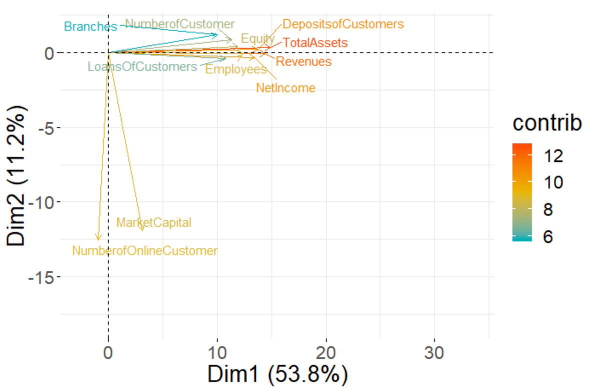


Figure 7.5: Bank-size Variables plotted on the two largest variance explained components - NB Database

The number of online customers points to the other direction and is thus negative correlated. Furthermore, it is also visible that the number of online customers and market capital are the main contributors of the second components, whereas the other variables contribute to the first component. Those two principal components explain clearly different information about the dataset.

Note that the number of online customers is due to the lack of information computed by the number of customers times the percentage of online banking customers in a particular country. Also, market capitalisation is described as an unstable measure in Chapter 4. Adding this component would not add information on the bank-size but mostly about the rate of online customers in a particular country. If the goal is to measure the availability of online banking users in a country, it is recommended to include this separably in the model. The second principal component will thus not contribute to measuring bank-size and is therefore excluded from the analysis. Only the first principal component is used to reconstruct bank-size: it explains 52 percent of the data variance. This seems low but it is the most optimal function of bank-size explaining the most variance as possible.

As specified, the first principal components is a linear combination of the original bank-size variables weighted by their contribution to explaining the variance of the dataset. The exact contributions of each of the variables towards the two principles can be found in Table G.1 (Appendix F). The equations of the bank-size component are shown in equation 7.1 and 7.2. Equation 7.1 is the bank-size component for the Logistic Analysis and Equation 7.2 for the Negative Binomial Analysis. It is visible that the contribution of each bank-size measure to the total bank size component does not differ very much. Both equations are used to calculate the new "bank-size" variable that will be analysed in the next chapter.

$$\begin{aligned}\textbf{BankSize} = {} & 0.321 \times Revenues \\ & +0.321 \times Equity \\ & +0.399 \times TotalAssets \\ & +0.079 \times MarketCapital \\ & +0.372 \times NetIncome \\ & +0.290 \times NumberofCustomers \\ & +0.334 \times Employees \\ & +0.273 \times Branches \\ & +0.294 \times LoansofCustomers \\ & +0.363 \times DepositsofCustomers \\ & -0.022 \times NumberofOnlineCustomers\end{aligned} \tag{7.1}$$

$$\begin{aligned}\textbf{BankSize} = {} & 0.379 \times Revenues \\ & +0.313 \times Equity \\ & +0.391 \times TotalAssets \\ & +0.081 \times MarketCapital \\ & +0.354 \times NetIncome \\ & +0.298 \times NumberofCustomers \\ & +0.324 \times Employees \\ & +0.262 \times Branches \\ & +0.283 \times LoansofCustomers \\ & +0.361 \times DepositsofCustomers \\ & -0.025 \times NumberofOnlineCustomers\end{aligned} \tag{7.2}$$

7.3. The Logistic Model

Logistic Regression is a useful technique to get a deeper understanding of the impact of several independent variables on a single dichotomous outcome variable. The dependent variable in this model is the binary variable: "is targeted".

7.3.1. Dealing with an unbalanced dataset

Some researchers argue that there is no benefit in creating a balanced dataset for Logistic Regression, it is only important to use all available data (Crone & Finlay, 2012). For this case, the Logistic Regression did not perform very well and it was expected that this was caused by the unequal distribution of the "being targeted" variable. The number of banks targeted in the dataset is 12188, whereas the number of not targeted banks is 403. The reason for the unequal dataset is twofold. Firstly, the constructed dataset is merged with Natalius (2018), and the targeted banks did match way more often compared to non-targeted banks. Secondly, targeted banks can appear multiple times in the dataset when they are targeted by different malware and/or in multiple years. There are 877 unique banks being targeted versus 403 not-targeted banks.

To deal with an unbalanced dataset the minority class was over-sampled using the Synthetic Minority Over Sampling Technique (SMOTE) method (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). This method assures that the minority class, non-targeted banks, became equally distributed among the targeted banks. For this, two types of SMOTE-functions have been used; one to over-sample the categorical variables and one to over-sample the continuous variables. This method is applied using the "imbalanced-learn" python packages (Appendix C). The new dataset contains 12188 non-targeted and targeted banks and keeps all the data as (Crone & Finlay, 2012) suggested. The performance of both logistic regression models are examined on their performance. For this purpose, two performance measures are used: ROC-curve and the McFadden R-square. The results can be found in Appendix H.1 and are discussed below.

The **Receiver Operating Characteristic (ROC)** curve shows the ability of the binary classifier Swets (1988). In this case, the true positives are plotted against the false positives. The

Area Under the Curve (AUC) is a metric to asses the performance of the ROC curve (Chawla et al., 2002). A value between 0.7-.8 refers to good, 0.8-0.9 to very good, and a value over 0.9 is even an excellent predictor (Šimundić, 2009).The AUC-value of the unbalanced dataset is 0.73 and the adjusted dataset using the SMOTE method has a value of 0.82.

The **McFadden R-square** is a 'goodness of fit' measure determining the predictive power of the model. The R-square calculates how strong the relationship is between the explanatory variables and the outcome variables. A McFadden R-square value between 0.2 and 0.4 represents an excellent fit (McFadden et al., 1977, p. 35). The McFadden R-square of the unbalanced dataset is 0.14 which is below the standard. The McFadden R-square of the balanced dataset is 0.34 and represents an excellent fit.

Both the AUC-value and the McFadden R-square measure show a significant increase in the performance of the model. For this reason, the analysis will be continued with the balanced dataset adjusted by SMOTE method.

7.3.2. Performance of logistic model

Prior to presenting the findings of the regression model, the model itself has to be checked on its reliability as well. The performance of the Logistic Model is examined to specify if valid conclusions can be drawn from the model. The performance test exists of a range of examinations. The model will be evaluate on the fit with the model using deviance statistic and the McFadden R-Square. Moreover, the model will be evaluated on accuracy and the ROC curve. The information of the performance of each step of the Logistic Model can be found at the bottom of Table 7.1. Here, only the performance of the complete Logistic Model are discussed, which is the last column of Table 7.1.

The ROC curve and the McFadden R-square are already explained and discussed in previous section and it is acknowledged that the model preforms very well.

Deviance statistics measure the 'goodness of fit' of the generalised linear model. Two types have been reported by the logistic analysis: the null deviance and the residual deviance. The null deviance shows how the response variable is predicted by a model that includes the intercept. Intercept is the expected mean value of 'being targeted' when all X-variables are equal to zero. The difference between the null and residual deviance is 11353.05 with the associated p-value of 0. The p-value is lower than 0.01, meaning that the null hypothesis - the coefficients of the added variables are 0, can be rejected (Singer, Willett, Willett, et al., 2003). This means that the logistic model is improved by adding explanatory variables.

Besides the model itself, it is also examined on its predictive **accuracy**. In other words, how much is correctly predicted by the model. The correctness of the prediction is done by calculating the percentage of true-positive and false-negative over the total positive and negative classes (Chawla et al., 2002). The correct categorisation of the logistic model is 77 percent.

Endogenity in the Logistic Model

Besides the performance measures, it is important to investigate if the data is over-dispersed, meaning that there is "unobserved heterogeneity in terms of a missing structural factor that leads to concentrations of observable events" (Tajalizadehkhoob, Böhme, Ganán, Korczyński, & Eeten, 2018). It is expected that endogenity caused by unobserved heterogenity plays a part in the regression models. Endogenity means that the explanatory variables correlated with the error term. The error term is everything that cannot be explained by the independent variables of the model. In the case of endogenity, the unobserverable error term is associated with the observed independent variables, also referred to as unobserved heterogenity. In this model, it might be a possibility that similar reasons triggering criminals to target the bank is also predictive for banks having two-factor authentication. ”Rich” banks are able to spend money on two-factor authentication to protect their assets. Concluding, the error-term contains information about the target frequency that is correlated with the independent variables. The implication of this could generate biased estimates. However, for binary outcomes, over-dispersion can only be meaningful measured if the set of variables with identical outcomes can be grouped. In the dataset there are 4 variables continuous measured, thus grouping is not possible.

7.3.3. Results of the Logistics Model

The goal of the Logistic Regression is to explain the probability of a bank being targeted. Because of the expected correlation between various independent variables, it is chosen to preform a step-wise logistic analysis in order to identify how variables change when others are added. An overview of the results can be found in Appendix H.

Table 7.1: Summary of Stepwise Logistic model

	Dependent variable:				
	Stepwise				
	(1)	(2)	(3)	(4)	(5)
banksize	0.535***	0.504***	0.473***	0.376***	−0.055*
	(0.024)	(0.024)	(0.024)	(0.024)	(0.025)
auth2FA		3.003***	3.273***	3.308***	4.662***
		(0.175)	(0.185)	(0.188)	(0.242)
langEnglish			0.814***	0.800***	0.676***
			(0.089)	(0.094)	(0.119)
langFrench			0.694**	0.880***	0.756**
			(0.227)	(0.235)	(0.280)
langDutch			1.311***	1.310***	1.838***
			(0.197)	(0.201)	(0.266)
langPortugese			4.306***	4.096***	5.061***
			(1.104)	(1.062)	(1.078)
langGreek			3.954***	4.708***	4.275***
			(1.092)	(1.107)	(1.099)
pop_score					5.337***
					(0.139)
has_parent				1.890***	1.205***
				(0.101)	(0.144)
BrandValue				5.489***	2.287***
				(0.517)	(0.570)
Observations	24,376	24,376	24,376	24,376	24,376
Log Likelihood	−13,789.500	−13,512.610	−13,203.350	−12,831.780	−11,219.630
Akaike Inf. Crit.	27,638.990	27,089.230	26,508.710	25,769.550	22,547.260
Deviance Statistic	6213.319	6767.084	7385.606	8128.76	11353.05
P-Value	0	0	0	0	0
McFadden R-square	0.1838678	0.2002551	0.2185588	0.2405506	0.3359655
Accuracy	0.6098211	0.6170414	0.6320971	0.7077043	0.7713735
AUC	0.733591	0.7447397	0.7574805	0.7661617	0.8180442
Note:					*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets

Bank-size

The aim of this research is to explore if bank-size measures influence target selection. The results in Table 7.1, prove that bank-size is a significant predictor for ”being targeted”. Noteworthy, in the four three models bank-size has a small, but positive influence on ”being targeted”. In the last model, bank-size has a negative effect. This controversy is presumably created by the interaction of the independent variables. For example, bank-size does correlate with domain popularity and the French language. Subsequently, domain popularity relates to being part of a banking group. This makes it too complex to adequately define the needed interaction variables, as is explained in previous section. Since bank-size does have a positive effect on ”being targeted” without domain popularity added to the model, and it is known from the descriptive analysis that large banks are being more targeted, the assumption is that bank-size has an weak positive significant effect on ”being targeted”.

Domain popularity, brand value and being part of a banking group

According to the results in Table 7.1, domain popularity is the strongest predictor whether a bank has been targeted. A bank belonging to a banking group or banks ranked in the top 500 brand values are also likely to be targeted. Here, brand value is even a better predictor compared to being part of a banking group.

Language-use on the website

It is expected that the largely spoken languages would have a significant effect on ”being targeted”. However, the results in Table 7.1 show that various - not largely spoken languages have an effect on ”being targeted”. Especially Portuguese and Greek are very important predictors. For this reason, the significance of the languages have probably other potential explanations. These potential explanations are summarised below:

- English is related to banks that are part of a banking group (Figure E.11). It is expected that these banks operate internationally and, therefore, offer the English language on their website. Since ”being part of a banking group” is even more significant towards target selection, it has a higher coefficient. It is expected that the underlying reason for the significance of the positively correlated English language is caused by the banks international orientation instead of the ease of attack.
- The influence of the French and Dutch language can be explained by three potential causes that are based on the descriptive analysis. First, it can relate to the bank-size, since both countries have large banks. Second, it can relate to the Belgium country, where banks have often two-factor authentication, increasingly being targeted. Third, French banks correlate with being part of a banking group, thus they operate internationally and have large assets. This could also make the more attractive for crime.
- The influence of the Portuguese and Greek language can be explained by the fact that both language have a high equity, thus they are also larger banks.

Two-factor authentication

The two-factor authentication has a positive effect on the chance of ”being targeted”, meaning that a bank with a two-factor authentication has a higher probability of being targeted. This is interesting since this security measure is supposed to make online banking safer, not more vulnerable. Two explanations could underlie this positive relation. First, two-factor authentication is not a barrier for adversaries, and thus other criteria are more important to consider for cybercriminals. Experts in this research, as well as the research of Natalius (2018), claim that it is relatively easy for criminals to bypass the two-factor authentication. Many banks may have implemented this security, also due to legislation, but it seems not to be a serious obstacle for professional criminals anymore. The second explanation concerns the correlation with two-factor authentication and other independent variables. For example, various languages highly correlate with two-factor authentication. The collinearity between language and two-factor authentication can change the coefficient of the two-factor authentication, and therefore, the Logistic Model shows a positive correlation coefficient.

7.4. Negative Binomial Regression Model

The Negative Binomial Regression Model assumes that the dependent variable is an observed count, like the number of targets. In multiple regression, there are various independent variables in the model. This allows fitting a more sophisticated model with several variables explaining ”target frequency”. In multiple regression, we still estimate a linear equation which can be used for prediction. The interpretation of the coefficient in the Regression Model is the effect of a unit change in the independent variable on the dependent variable, holding constant all other independent variables in the model.

The Poisson model also has as dependent variable an observed count, but this model has the restrictive assumption that the variance and mean should be equal like the poisson distribution. The data shows that the mean (10.27) is smaller than the variance (1101.37) and, therefore, a Negative Binomial distribution is used. (Hilbe, 2011).

7.4.1. Performance of the Negative Binomial Model

To evaluate the performance of the Negative Binomial Model, three performance measures are used: the McFadden R-square, deviance statistic, and the dispersion. These performance measure can be found at the bottom of 7.2, only the performance of the complete negative binomial will be discussed (last column).

Similar to the Logistic Model, the **deviance** statistic and **McFadden R-square** are used as performance measures. The explanation will not be repeated, only the outcomes of those tests will be described. The deviance statistics, thus the difference between the null and residual deviance model is 35969.12. The associated p-value is 0, which is lower than than 0.01, meaning the null-hypothesis can be rejected. The Negative Binomial Model with the explanatory variables is significant better than the null model. The **McFadden R-square** is 0.22 representing a strong enough predictor.

The Negative Binomial Model shows a **dispersion parameter** of 10.29, the dispersion parameters are significantly higher than 1, indicating that the data is over-dispersed, meaning that there is ”unobserved heterogeneity in terms of a missing structural factor that leads to concentrations of observable events”. As aforementioned, the endogenity was already expected. Conclusions should be made carefully, since the implication of over-dispersion is that it could generate biased estimates.

7.4.2. Results of the Negative Binomial Regression

Unique count, selected in Chapter 6, is used to evaluate the results of the Negative Binomial Metric. This seems to be the best metric to count the number of targets since it is based on a legitimate explanation of how malware data is implemented. The formula and results can be found in Appendix H. This section discusses the most important results.

Bank-size

Bank-size has a positive influence on target frequency. The predictive value is not very high. Yet, in this case, different from the Logistic Model, bank-size is a better predictor for the frequency compared to the visibility related factors.

Domain popularity, brand value, and being part of a banking group

For domain popularity, the rank of its brand value, and being part of a banking group are

Table 7.2: Summary of stepwise Negative Binomial model

	Dependent variable:			
	stepwise			
	(1)	(2)	(3)	(4)
Bank-size	0.041^{***}	0.041^{***}	0.044^{***}	0.029^{***}
	(0.003)	(0.003)	(0.003)	(0.003)
auth2FATrue		-0.146^{***}	-0.153^{***}	-0.144^{***}
		(0.041)	(0.041)	(0.041)
langEnglishTrue			0.167^{***}	0.181^{***}
			(0.040)	(0.040)
langGermanTrue			−0.021	−0.034
			(0.055)	(0.055)
langFrenchTrue			0.178^{*}	0.200^{**}
			(0.077)	(0.077)
langDutchTrue			0.713^{***}	0.713^{***}
			(0.102)	(0.100)
langItalianTrue			-0.314^{**}	-0.266^{*}
			(0.107)	(0.106)
langSpanishTrue			0.875^{***}	0.810^{***}
			(0.098)	(0.098)
langPortugeseTrue			-0.955^{***}	-0.903^{***}
			(0.120)	(0.120)
langPolishTrue			0.352^{*}	0.477^{**}
			(0.172)	(0.173)
langHungarianTrue			0.370^{***}	0.457^{***}
			(0.110)	(0.109)
langSwedishTrue			0.176	0.364^{**}
			(0.124)	(0.124)
langLithuanianTrue			−0.787	−0.958
			(1.183)	(1.180)
pop_score				0.00000^{***}
				(0.00000)
BrandValue				0.001^{***}
				(0.0002)
Part_of_BankingGroup				0.201^{***}
				(0.030)
Observations	12,228	12,228	12,228	12,228
Log Likelihood	−32,374.190	−32,365.130	−32,211.210	−32,114.540
θ	2.585^{***} (0.042)	2.591^{***} (0.042)	2.686^{***} (0.044)	2.743^{***} (0.045)
Akaike Inf. Crit.	64,862.370	64,848.250	64,574.410	64,387.070
Deviance statistic	33845.81	33927.71	35200.51	35969.12
Associated p-value	0	0	0	0
McFadden R-square	0.2103139	0.2105349	0.2142895	0.2166476
Dispersion	12.9139	12.74849	11.03169	10.28711
Note:				*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets

all positive correlated with the target frequency. ”Being part of a banking group” is a strong predictor compared to brand value and domain popularity.

Language-use on banks website

The language-use is after two-factor authentication the strongest predictor. Similar to the Logistic Model, many languages show their importance in the Negative Binomial Model and have potential underlying reasons causing this significance. A potential explanation for each language is described below. These insight are mainly based on Figure E.11 from the descriptive analysis.

- The Spanish and Portuguese language are connected with the French country. This means that French banks often offer the Spanish and Portuguese language. The French banks are the larger banks. The importance of the Spanish and Portuguese language is thus probably caused by its relation with larger banks. Moreover, Spanish and Portuguese are correlated (0.5). Interaction terms should be added to identify if the Portuguese language indeed negatively correlates. The descriptive analysis showed that the Portuguese language increases the target frequency instead of decreases.

- The Italian language is highly related to domain popularity. Similar to the relation between Spanish and Portuguese language, the interaction with the Italian language and domain popularity will probably result in a negative coefficient.

- French relates to the Dutch language and since banks offering those language are also the larger banks, the size is probably the explanation of its positive influence on the target frequency.

- The Polish language scores high on visibility compared to others, and is probably related.

- The Swedish language scores high on having two-factors authentication. Adding interaction will probably change the positive coefficient to a negative coefficient.

- The Hungarian language correlates relative highly with various languages, such as English. The interaction between those language will probably have an effect on the model.

To check if the above hypotheses are true, an extra Negative Binomial Model is build. In this model, interaction terms are added aligned with above statement. Although this model has interaction variables to show the relationships, it is not complete and therefore, the model will not be used to interpret the significance of the independent. Still, the model provides insight into the relations of some language factors with domain popularity and size and it validates above statement.

The following interaction variables are added: two-factor authentication and the Swedish language, the Hungarian language and the number of languages and the English language, the Italian language and domain popularity, the Dutch and French language, the Polish language and domain popularity and, finally, the Spanish and Portuguese language. All of the above mentioned interaction terms are significant, meaning that those variables indeed interact with each other. The result can be found Table H.4 (Appendix H). From these results can be concluded that Portuguese and Swedish are after adding interaction terms not significant anymore, the Italian language is positive instead of negative, the Polish language is negative instead of positive, and The Hungarian, French and Dutch language increase in prediction capacity. This model confirms that interaction indeed has a big impact on the interpretation of the correlation coefficients. It also confirms, as suspected, the importance of language-use is probably caused by bank-size or popularity.

Two-factor Authentication

Having two-factor authentication correlates negatively with the target frequency. This is an expected result since it would be more difficult for adversaries to target. Still, this relation can be clarified by two potential explanations. First, two-factor authentication is indeed effective and it is difficult for criminals to successfully target such banks. Adversaries try to target banks with two-factor authentication, probably because they are larger (Chapter 6), and will move on to other banks when they cannot succeed. Secondly, banks with two-factor authentication will be more security mature i.e. they will have better rule-detection systems. From the descriptive analysis it is known that two-factor authentication correlates largely with certain languages i.e. Swedish and Danish, and thus also countries that are cybersecurity mature. Two-factor authentication also is connected with the French and Dutch language which are the banks with the largest bank-size. The collinearity between two-factor authentication and the other independent variable or other underlying factors i.e. cybersecurity maturity affects the target frequency.

7.5. Additional information from analysis

The influence of country on target selection

Comparing the results of the Logistic Model with the Negative Binomial Model, it is remarkable that countries prove to have a significant influence towards the frequency of targets, but not IF a banks is being targeted. A hypothesis could be that cybercriminals do not prefer certain countries above others. Yet, when a certain bank in a particular country seems successful, cybercriminals try to also target the other banks in the country. This will be further discussed with the experts in the next section.

Regression Model including bank-size measurement based on feature selection

As mentioned, feature selection is applied for selecting the most important bank-size measurements and to reduce collinearity. This is all explained in the Appendix F. The selected bank-size measures from the feature selection are analysed by both regression models in Appendix H. The results show that Equity and Total Assets were significant factors influencing whether a bank is being targeted. Total Assets and Number of Customers influenced the target frequency. When adding single bank-size measures those seems to be the best predictors, whereas adding the bank-size component, domain popularity is a better predictor. Other than that, similar results as the regression models described above came forward.

Verifying literature theories

In Table 3.1, a few theories from the literature study have been identified and they will be tested in this section.

1 Banks with less than 250 employees will more likely be targeted by Banking Trojans (Symantec, 2015).

2 Organisations with more than 500 employees will likely experience a DDoS attack (Arbor Networks, 2016)

3 Revenues under 35$ have a higher risk and are more vulnerable to being targeted Beazley Breach (2016).

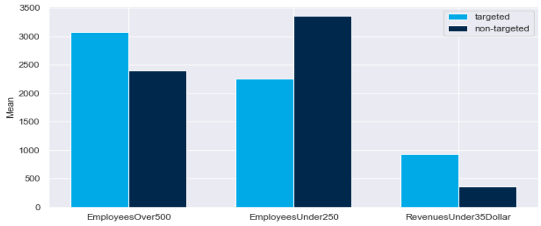


Figure 7.6: Barplot of tested hypothesis

Independent Variable	Estimate	Pr(>|z|)
Employees under 250	-1.09248	2e-16 ***
Employees over 500	1.07529	2e-16 ***
Revenues under 35$	1.43290	2e-16 ***

*p<0.1; **p<0.05; ***p<0.01

Table 7.3: Result: Logistic Regression verifying/ falsifying theories

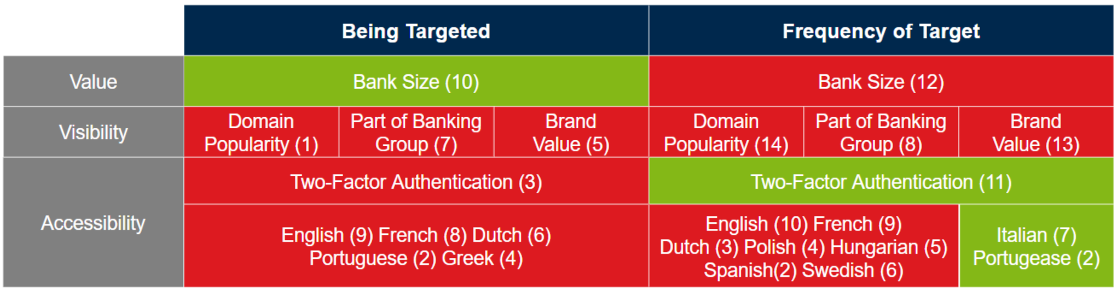
A Logistic Regression is used to predict the probability of the conditions (employees under 250, employees over 500 and revenues under 35) to fall into ”being targeted” or ”not being targeted”. First, dummy variables will be created for every condition. Then, the dummy variables will be included in the Logistic Model. The outcomes can be found in the Table 7.3 and visualised in Figure 7.6.

Two out of three statements can be verified. However, the confirmed statements seem to contradict each other: banks with more than 500 employees are more likely to be targeted, and banks with revenues under 35 $ have a higher risk of getting targeted. This can be explained using the principal component analysis plot in Figure 7.4. Here is visible that revenues has a negative effect, pointing downwards, compared to employees. The reason for revenues being negative needs further research, but is probably a result of errors in the data collection part, which is already acknowledges in chapter 4.

Summary - Chapter 7

This chapter established the regression model and dealt with the problem of multicollinearity by performing a Principal Component Analyse.

The results of the regression are summarised in the figure below. Here, the green variables result in decreasing (the number) of being targeted and red results in increasing (the number) of being targeted. Furthermore, the variables are ranked on their predictability. Noteworthy, it is known that endogeneity occurs in the model and that the banks' characteristics are highly correlated. These properties might affect the estimated coefficients of the regression models.



	Being Targeted			Frequency of Target		
Value	🟩 Bank Size (10)			🟥 Bank Size (12)		
Visibility	🟥 Domain Popularity (1)	🟥 Part of Banking Group (7)	🟥 Brand Value (5)	🟥 Domain Popularity (14)	🟥 Part of Banking Group (8)	🟥 Brand Value (13)
Accessibility	🟥 Two-Factor Authentication (3)			🟩 Two-Factor Authentication (11)		
	🟥 English (9) French (8) Dutch (6) Portuguese (2) Greek (4)			🟥 English (10) French (9) Dutch (3) Polish (4) Hungarian (5) Spanish(2) Swedish (6)		🟩 Italian (7) Portugease (2)

Figure 7.7: Visualisation of the results

8
Expert Interviews

The previous chapter presented the results of the regression analysis. This chapter validates these results by conducting expert interviews. The interviews review the regression model and seek to obtain critical opinions on the key characteristics influencing of target selection. It opens up a possibility to discuss unidentified potential factors for target selection and provides a broad view on the topic.

8.1. Approach

The conducted interviews follow a semi-structure interview that combines a pre-determined set of questions with questions that came up during the interview. This gives the opportunity to explore a particular response or theme. The interview protocol is send a week before the day of the interview to the experts. This gives them the possibility to look at the questions and understand what can be expected. After the interview, the elaborated script (Appendix I) is send to the experts to confirm the completeness and accuracy of the script.

8.2. Interview protocol

The interviews consist of two phases. The first phase aims on obtaining an experts perspective on potential factors that affect target selection. The second phase focuses on the experts opinion and professional interpretation of the developed model, including some other results of this thesis. The interview protocol and the questions can be found in Appendix I.

8.3. Profile of experts

The interviews are chosen with the objective to discuss topics with experts with diverse specialisations and perceptions. Experts from two types of banks were willing to participate; a national focused Dutch bank and a large Dutch bank. Also, an expert from the cyber security organisation ThreatFabric was invited, and his view was important to avoid biases and to balance the interviews in the banking sector. For security and privacy reasons, the names of the experts will not be published. This chapter provides a summary of the interview outcomes.

A Expert Mobile Threats at ThreatFabric. ThreatFabric is a cyber security company supporting the financial sector to pro-actively detect known and unknown mobile threats.

B Security Advisor and Risk Lead at ABN AMRO.

C Security Analyst and a Senior Advisor Criminal Risk Management at a Dutch national bank.

8.4. Interview analysis

The goal of the interviews is to gain insight in the practical expertise related to target selection and to validate the results. The validity and the results will first be discussed followed by additional insights that are never mentioned in previous research.

8.4.1. Validation of results

Experts are asked questions concerning the relevance of this research, if the results reflect their expectations and how the model could be improved. The script of the interviews can be found in Appendix I.1 and will be summarised in this paragraph.

Value of research

The both experts from the financial sector, agreed that the model will provide added value in understanding the underlying factors of target selection. It supports decision-making in the prioritisation of implementation of security measures.

Model validation

The validity of the results and the model are described below.

- **Bank-size**
  Experts of both banks confirmed that bank-size, in combination with brand-awareness, plays an important role in target selection. It is assumed that big banks play a more important role in target selection, whereas the number of clients and their deposits were primarily mentioned as main influence representing bank-size. Another argued that adversaries are simply too lazy to conduct a detailed reconnaissance phase or to gather intelligence on the bank-size.

- **Language**
  When language was discussed as an important factor it was suggested that this is most probably related to the ease of making an fake-webpage to harvest credentials. Yet, the experts confirmed that they have seen those fake web-page in perfectly written Dutch. In reality the used language may have lost a bit of its importance.

- **Domain popularity**
  The significance of domain popularity was not immediately recognised by the experts as an important factor that influences target selection. In most cases, the experts could understand the importance of this factor, but they were still surprised by the results.

- **Two-factor Authentication**
  The experts consider the Two-factor authentication as very important. Showing the results of the model the experts understood that the challenge for professional adversaries is not an overly complicated issue.

Model improvement

Three recommendations are summarised to further improve the model. First: the expert of the national bank expected that target selection is very much connected with the brand-awareness and it is therefore recommended to add it to the model. This is done by using a brand-value ranking (see Chapter 7). Moreover, the expert state that the risk of being targeted increases when a bank is operating in an international environment. The same applies when it provides different services i.e. private, investment, business banking. Being part of a banking group could be related to banks that operate on an international level, or when it provides multiple services. Being part of a banking group is therefore added to the model to identify if potential variables i.e. international operation or providing multi services might influence target frequency. A second missing factor is the ease to launder money. And finally, the large Dutch bank sees mobile malware as a potential future attack-vector and recommends to add data on mobile malware.

8.5. Key insights

From the interviews it is clear that most experts mention similar factors explaining target selection by banking malware; cost-efficiency and profitability influence the adversaries behaviour. The adversary will always (re-)act in order to ensure profitability. As long as the revenues remain higher than the operating cost, the cyber criminal will keep on operating, else other options will be exploited.

Additional insights are structured by the four aspects of defining a suitable target: value, portability, visibility, and accessibility (RAT in Section 1.2.1).

Interesting is that the insights related to operating at national level and the ease of money laundering are identified by several experts. However, their argumentation differs.

Value related characteristics

Operating at National or International Level; a factor that is never mentioned in other studies, but recognised by two experts, is that operating on a national level decreases the chance of being targeted. This has three reasons. First, operating on a national level decreases the chance that criminals in other countries know the existence of the bank. The barrier of the language is also more difficult to overcome when only the domestic language is used on the website. The other reason is the fraud detection of transactions. When a bank is focused on a national level, international transactions happen less frequent compared to international operating banks, thus it would be easier to detect fraudulent transactions. Moreover, operating at an international level could influence target frequency due to a more visible bank profile.

Targeting smaller banks; is recognised by the Fox-it analyst who monitors this behaviour. Smaller banks lack finances to have proper rule engines in place. On the other hand it is acknowledged that smaller banks do cooperate together and collectively outsource their security. Another experts has identified that adversaries first try large banks in a country. When they are successful they also try to attack smaller banks of the same country. It spreads like a virus.

Visibility related characteristics

Service provision by banks that provide multiple services, i.e. private, corporate and investment, have a higher risk to be attacked. Once adversaries saw that the attacks on customer services were successful, they moved on with attacking the investment or private branch of the same bank. Providing multiple services shows that there is enough money which makes a bank more vulnerable to banking malware attacks.

Prominent search results: Adversaries can be considered as lazy and they will likely just search on Google for big banks in a specific country in order to find suitable targets.

Accessibility related characteristics

Two-factor authentication and rule-based detecting; one of the threat analyst experts identified that the two-factor authentication was not very popular among the banks since it is easy to bypass. When the banks moved away from the two-factor authentication and implemented rule-based engines or intelligence providers, the banks were able to actually stop the attacks. This pattern is confirmed by the both bank experts. Proper rule-based detection and transaction monitoring systems enormously reduce the number of successful attacks. Before those countermeasures were in place, millions of euros left the bank on a weekly basis. The success-ratio of the attacks lowered and adversaries moved over to another tactic.

Pattern of adding/removing banks from configuration file. Banks are sometimes removed from the configuration list and are added again after a certain period of time. It is likely that cybercriminals first try to target particular banks such as large banks or banks for which they already have recruited money mules for. When they are not successful they delete the bank name and they move on to attack other banks. However, after updating their malware and successful circumventing the security of the bank of interest, they again listed these banks in the configuration file. With the Dyre malware, a 'shit list' was created by the adversaries. This list listed all the banks that were able to detect the malware resulting in non-successful attacks. Those banks were removed from the configuration file.

Portability related characteristics

The Ease of money laundering is often mentioned as a important variable by the experts. There are three causes. Firstly, the social condition of a country; e.g. poverty and employment will influence if people are more willing to function as a money mule. Underprivileged communities or socio-economic underclasses are easily and successfully recruited. The second cause is the so-called "onboarding of digital customers". Banks that enable online registration for a bank account are more vulnerable for the creation of money mule accounts. This also applies for allowing multiple accounts for one person, which is actually meant to support freelancers but is also a vulnerability for money mule account.

Impact of politics and technological development of a country

The influence of political factors is already mentioned in other researches, i.e. money transfer policies of the country and ICT development index (see Table 1.1). It is confirmed by the experts that proper transaction monitoring systems massively decreased the target frequency. European banks are legally obligated to pay attention to unusual transaction patterns and to transactions that by their nature involve a higher risk of money laundering or terrorist financing. Legislation in the Netherlands play a role in maturing the banks security measures.

8.6. Updated list of characteristics influencing target selection

The experts mentioned some variables that are not mentioned in earlier studies. This expands the current list (Table 1.1), the banks' characteristics are structured according to the RAT theory.

- **Value**: bank-size, and at country level: adaption rate of online banking, financial status (GDP), ICT development index, rate of banking penetration, maturity of online banking system.

- **Visibility**: brand popularity, domain name visibility, banks' attack record, website domain popularity, ownership of the bank, operating international/national, prominent search result.

- **Accessibility**: degree of cooperation between financial institutions and law enforcement within a country, but also the degree of cooperation between financial institutions, 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, language of the online banking, maturity of online banking system.

- **Portability**: the ease to recruit money mules (on-boarding of digital customers, allowing multiple accounts per person). At country level: ease of recruiting money mules related to social factors and awareness of customers, money transfer policies of the country

Summary - Chapter 8

This section aims to get a critical opinion on target selection from experts and to validate the results from the regression analysis in previous chapter. Due to the expert interviews, some factors are revealed that never have been mentioned in other research. These factors are: type of service (national vs international and number of branches - investment, private and public), google hit score, on-boarding of digital customers, and the number of accounts per customer.

Besides, the results from the previous chapter are validated by experts. Bank-size, language, and two-factor authentication are recognised by experts as important factors that influence the target selection. They were surprised about the effects of the domain popularity, but also by the controversial results of two-factor authentication. Interesting was that experts sometimes had different explanations why a factor may influence the target selection: e.g. the ease to recruit money mules or operating on international versus national level.

The next chapter identifies the limitations of this research and reflects on the potential impact on the outcomes of the regression model.

9 Discussion

This research acknowledged several limitations. This section addresses these limitations and describes potential implications for the results. In addition, this chapter reflects on the presented results in the previous chapter by making a comparison with earlier target selection studies. Finally, this chapter ends by reflecting on the suitability of the RAT theory.

9.1. Limitations of this study

The limitations focus on two aspects: 1. data completeness and quality and 2. the model input.

9.1.1. Incomplete and inaccurate data

The most challenging part of this research was to collect the data, and process it in such a manner that is it reliable for the analysis. Complete and qualitative data is essential for the regression analysis. The limitations of the data are already discussed in Section 4.3.3 and 4.4.2. The potential impact will now be discussed and summarised.

- **Bank-size database**. As a consequence of the language barrier and the inconsistent use of definitions, the dataset is incomplete. Moreover, the data is for the larger part manually acquired ensuing (human) errors in the dataset. This potentially influences the data *accuracy*.
- **Target database**. Natalius (2018) has built a mechanism to match the banks URLs (those are listed in the configuration file) with the banks. He recommended building better mechanisms to reduce false negatives. Due to time constraints, this has not been improved in this research. Furthermore, the database shows a small gap in the data, which points towards *incomplete* data.
- **Other banks' characteristics**. This dataset relies on the data from previous researches.These should be reliable sources and it definitely saves valuable time, but it is never 100 percent sure that the data can be trusted. There was no time to validate the correctness of the data or to improve it.

The merging of the bank-size database with other banks' characteristics resulted in a considerable amount of data loss: the bank-size database had 2248 banks, but after the merging, only 1666 banks were left for the analysis.

The effect of above-summarised data limitations influence the quality of the data, and thus could lead to misleading results. To give an example: in Appendix H.1 it was found that bank-size did not correlate with the under performed unbalanced data, but it is significant in the balanced dataset. In data-science this phenomena of unqualified data on the results is also described as ”garbage in, is garbage out”.

Although the data quality is not perfect, the data is collected from multiple sources and is even cross-validated when possible, This research was able to collect data on a high number of bank-size measurements and a high number of banks, this is never done before and it can be stated that it is the best possible dataset.

Inadequate measurements

Brand value and *being part of a banking group* can be considered as not adequate measurements. Brand Value is based on a web-source that is arguably reliable. And, the determination if a bank is part of a banking group depends on the mechanisms matching URLs with (sub)entities. This mechanism matches the URL with the bank-name and also if it is part of another bank institutions. As mentioned above, this URL matching mechanism should be improved and it is unknown if the information if a bank belongs to another group can be subtracted from the URL. Therefore it cannot be trusted that the target database has the complete information if a bank belongs to the banking group. Nonetheless, those measures are added to explore the relationship with target selection as potential variable for future research. Those measures serve the objective and show that it is worthy to add those (more adequately) measurement to the model in future research.

9.1.2. Limitations of model input

Two limitations arise when evaluating the input of the model.

Missing portability related factors

Undoubtedly, many variables are missing in the regression model and many could be added to get a more comprehensive view of target selection. One underexposed aspect of the model should be addressed in particular, namely *portability*. Portability is the only aspect of a ’suitable target’ from RAT that is not part of the explanatory model. Portability, such as the ease to recruit money mules or enabling the registration of online customers, is in current literature and research acknowledged as an important factor influencing target selection and should therefore be in incorporated in the model.

Measuring banks’ characteristics and external factors

The variables in the regression should be measurable, but many variables are difficult to quantify. For example: measuring the use of other security measures next to the two-factor authentication, such as rule-based detection systems. It could be measured in the number of rules that are written, but this does not tell anything about the effectiveness of the system. However, there are a variety of security metrics that van be used to measure security maturity, such as the: Information Security Maturity Model (Saleh, 2011).

Next to the difficulty of measuring banks’ characteristics, also *external factors* play a vital role in the target selection process. For example, the demand and supply of underground forums. Banking malware attacks exist of many steps and adversaries need many (technical) skilled specialists. Even advanced criminal groups purchase services from the underground market, such as kits or money mule services. The use of banking malware is therefore highly dependent on the dynamics of the supply and demand on the underground forums. An example: when kit malware was not updated, the activity of the malware enormously dropped. It is difficult to include such external factors in a regression model.

9.2. Reflection on the used approach

This section reflects on a the used approach, namely collecting, preparing and analysing the data.

Data cleaning: biases in data availability

In the data cleaning process, choices had to be made for the incomplete dataset. These choices potentially introduces biases and give a wrong indication of the balance of small versus large banks. Since this research examines bank-size in target selection, biases related to the imbalance towards small/large banks need to be identified.

To give an example: some banks, being part of a banking group, only provide information on group-level. When there was no information for the individual banks, the values of the group were collected. However, if information was available, then the values were reported on bank-level. This approach could lead to an inconsistent view of the proportion of small versus large banks. Nevertheless, these circumstances only apply for less than 10 banks, it will, therefore, have a minimal impact on the results.

In addition, in Chapter 4, the credit unions in Poland are merged into one group since they use the same online banking system. However, German banks also collaborate in their online banking systems, but these are not merged since it was unknown for which banks this applied. This can give a wrong impression on the proportion of small versus large banks. This dissimilarity in approaches is very limited does not have a severe impact on the results.

Bank-size metric

When adding bank-size to the regression model, the model showed inconsistent behaviour because all features correlated and provided redundant information. This research solved this issue by utilising Principal Component Analyse and derive one bank-size variable out of all ten bank-size measurements. Chapter 4 acknowledged that some measurements i.e. Revenues, Net Income and Market Capital are not very reliable. These measures are also used in computing ”banksize”. It would be better to only conduct a PCA on measures that are easy observably in order to reduce noise in the bank-size component. Still, the method used in this research is the most ’honest’ way. Moreover, this is the first research to review bank-size measurements and define a bank-size metric. There is no research to collect manually data of over 2200 banks to construct the bank-size metric.

Regression analysis

From this research is clear that there was a lot of independence between the bank characteristics influencing the estimated coefficients. Since regression models perform better with independent variables, it would have been better, to perform, next to bank-size, multiple PCAs to ensure independence in the regression models. For example, combine the visibility factors into one component. This feature reduction makes it possible to add interaction terms to the regression models. Reducing the independence and enabling interaction terms will ensure that the coefficient can be interpreted. However, due to time constrains it was not possible to perform multiple PCAs, but with the descriptive analysis and human interpretation, it was still possible to recognise interaction and draw conclusions from the model.

Moreover, linear regression suggests that the relationship between the dependent and independent variables can be expressed in a straight line. Criminals constantly change strategies and methods to maximise financial gains. They show bounded rational behaviour. One of the experts even called criminals lazy since they copy-cat configuration files. It could be doubtful that the relation between banks characteristics and target selection is linear. According to an expert in the research of Natalius (2018), the model is never able to explain target selection sufficiently, and there is always a need for a combination of quantitative and qualitative methods. Even though the complexity of this problem cannot be captured in regression models, it provides insight into key characteristics influencing target selection. The approach is in line with the quote of Box (1980) who said: ”All models are wrong, but some are useful”.

Interview methodology

As described in Section 8.1, the results of the interview are verified with the experts by sending the transcript to them with the question if they have remarks or corrections. A better approach would be to send the insights/conclusions to the expert and verify those with them. Now the presented results are based on the writers interpretation which can contain biase. Sending the conclusions to the experts would verify their interpretation,

9.3. Comparison with earlier target selection studies

This research aims to get a comprehensive view of target selection in banking malware. This section compares earlier target selection studies with this research. It identifies similarities and dissimilarities.

Similarities

Similar to the research of Van Moorsel (2016), the decreasing accessibility - authentication method, is acknowledged as a short term solution and will sooner or later be bypassed by sophisticated malware. Financial institutions will strive to implement new available authentication methods. This research takes a step further by emphasising that the quality and even the diversity of the authentication method influences the intensity of target selection and indirectly the quality of the online security of a specific bank.

Dissimilarities

The research of Van Moorsel (2016) state that due to the increasing awareness and the sophisticated defence measures, large banks become less interesting to target. The trend of the increased popularity to target small banks has never been verified by scientific research nor can it be confirmed with this research. This research identified in the descriptive analyse that smaller banks (related to the number of customers) have been targeted less. Also, in the Negative Binomial Regression was visible that the number of customers increased the target intensity over the year 2016 and 2017. In addition, it appears that banks with over 500 employees were more likely to be targeted. In the quantitative data analyse, there is no clear trend visible of targeting smaller banks. Also the qualitative analysis, the expert interview, identifies that cybercriminals not only chose the ”easiest” target, but that they also have their preferences for certain banks.

The experts in the research of Van Moorsel (2016) draw the link that smaller banks have lower security measures. However, smaller banks in Germany cooperate closely together and use third-party services to secure their transactions. Smaller banks do not by definition have lower security measures as long as they have options and means to mitigate security risks. When adversaries can conduct a successful attack on a large bank, they will continue the targeting actions as long as they remain profitable. According to the experts, adversaries first try to attack the large banks, and they switch to smaller banks when they become less successful.

Tajalizadehkhoob (2013) showed that English webpages are an attractive characteristic for target selection in the Zeus malware for the years before 2012. Also, Natalius (2018) highlights the attractiveness of the English language, which is also in line with the results of this research. However, this research and the research of Natalius (2018) show that multiple languages are relevant.

9.4. Reflection on RAT

The goal of this research is to identify if RAT is able to explain victimisation in digital crime. According to Leukfeldt and Yar (2016) digital victimisation shifts from ’value’ and ’visibility’ towards ’online accessibility’. This thesis advocates the importance of value, namely bank-size, and visibility, such as domain popularity, being part of a banking group, and brand value for target selection of banking malware. These variables influence target selection but are not only connected to online accessibility.

Summary - Chapter 9

This research acknowledged several limitations that might affect the results. First of all, the limitations related to the data can lead to misleading results in the regression model. Furthermore, portability and external factors, i.e. the demand and supply of the underground market, are recognised by experts as important variables influencing target selection but are not included in the regression model.

Improving the model can be done by performing multiple principal component analysis to combine banks characteristics and, thereby, reduce independence in the model. Another improvement would be to add control variables related to the social context of a country in order to eliminate false-correlations.

Comparing to earlier target selection studies, this study also acknowledges that decreasing accessibility, such as implementing two-factor authentication, is a short term solution. Nonetheless, it does not observe that larger banks become less interesting targets; the outcomes of the regression model showed that larger banks are still an attractive target. Also, offering largely spoken language on the banks' website does not ease the attack anymore.

It has been confirmed that RAT is a suitable theory for explaining victimisation in target selection research. The next chapter answers the primary research question and provides some recommendations founded on the gathered intelligence of this research.

10

Conclusions and recommendations

In the previous chapters, substantial effort is put into developing and analysing a regression model that is able to explain target selection. This chapter summarises the conclusions and recommendations of this research paper. This is done by providing a short recap of the research objective. Thereafter, the sub-research questions will be answered as they were introduced in Chapter 1. Subsequently, the information from these sub-research questions will be leveraged to answer the primary research questions. Finally, this chapter closes by proposing some recommendations founded on the outcomes of the analysis in this research

10.1. Recap research objective

The goal is to establish an explanatory model explaining target selection of banking malware. Current target selection studies analyse multiple banks' characteristics, but they were limited to one type of malware and they differ in geographical focus. Natalius (2018) was the first researcher to develop a model which analyses different types of malware. However, he didn't take financial and market features into account which are, according to experts in the field, essential in explaining target selection. For this reason, the goal of this research was to include bank-size measures in the model and consequently to explore the relationship with target frequency. In doing so, a comprehensive and insightful model has been created to explain target selection in banking malware.

10.2. The answers to sub-research questions

This section lists the four sub-research questions. The questions will be answered from the analysis and the outcomes in previous chapters.

1. How does the current banking malware threat landscape look like?

This question is very broad and, therefore, the STIX framework is used to structure the information comprehensively. The following three concepts were discussed to map out the threat landscape of banking malware.

- **Threat actors**
  The criminal ecosystem of banking malware mainly exists of crime facilitators, digital robbers and state-actors. They operate individually, in groups or in organisations. A banking malware attack needs various technical skilled specialists to support an attack and, therefore, cybercriminals work mostly in networks. The networks can be related to organised cybercrime groups or individuals who collaborate via purchasing/selling services on the underground forums. These forums provide services along the attack chain of banking malware. Even organised cybercrime groups make use of the underground forums to purchase exploit-kits or for money mule recruiting services.

- **Tactics, Techniques and Procedures**
  There are multiple techniques used by adversaries to infect and harvest credentials of online banking customers (Chapter 2). A typical banking malware scheme starts with the threat actor writing the malware. The adversary selects the banks from which he/she wants to target the customers by listing the bank domains in the configuration file. This file contains a list of URLs (bank domains) and provides information about the behaviour of the specific malware. When an infected malware user browses to their banking website, which matches the URL in the configuration file, the malware orders the malicious server to inject a modified page. The user inserts the credentials in the modified page, and the information is sent to the command and control server of the adversary.

- **Campaigns**
  Banking malware types differ in functionalities and geographical focus. Functionalities, versions and variants of malware evolve over the years and new malware enters the threat landscape. Nonetheless, older malware variants are updated and have remained active over years.

2. How is bank-size delineated in current literature and how can bank-size be measured?
In order to select suitable bank-size measures for this thesis, existing studies had to be analysed regarding how they measure bank-size. Research topics in the field of economics, cyber, and previous target selection studies are examined on how they measure bank-size. From these studies it is recognised that a standard metric of bank-size does not exist.

Considering the bank-size measurements in economics, cyber, and target selection studies in combination with the ranking of Schildbach and Schneider (2017), a selection was made of the bank-size measurements for this thesis. Thorough research on assessing the advantages and disadvantages of these variables was conducted to select the measurements. A total of eleven bank-size measurement were selected: revenues, total assets, equity, net income, risk-weighted assets, number of employees, branches, number of clients, number of (online) customers, loans and deposits of customers.

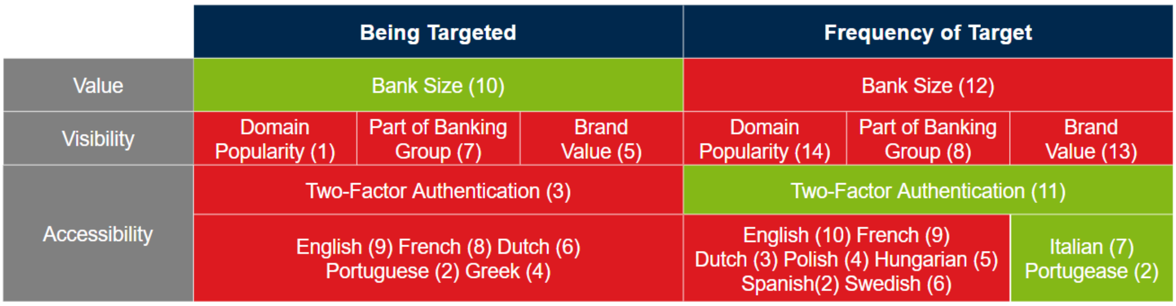
3. Does bank-size, online language-use, domain popularity, brand value, being part of a banking group and two-factor authentication influence if a bank is targeted or influence the frequency of targeting?

Before the regression model could be established to provide answers to both parts of the question, considerable effort is put into collecting, cleaning and processing the data. In addition, PCA was performed to derive one single bank-size variable from the eleven collected bank-size variables. This analysis deals with multicollinearity in the regression model. Below, the results of both models are listed.

Figure 10.1 shows the result of the regression analysis. The first column ’Being Targeted’ shows the relationship of the banks’ characteristics with the likelihood to be targeted. The second column ’Frequency of Target’ shows the relationship of the banks’ characteristics with the frequency to be targeted. The red visualised variables increase the probability to be targeted or being a more frequently target. The green variables lower the probability to being targeted or being less frequently targeted.

The results are summarised by the following explanations:

- **Bank-size**
  Bank-size decreases the probability of being targeted, but increases the probability to be more frequently targeted. The logistic model showed that smaller banks result in more often being targeted, which is controversial to what we have seen so far in the descriptive analysis. However, this negative coefficient is caused by the interaction with popularity



	Being Targeted			Frequency of Target		
Value	🟩 Bank Size (10)			🟥 Bank Size (12)		
Visibility	🟥 Domain Popularity (1)	🟥 Part of Banking Group (7)	🟥 Brand Value (5)	🟥 Domain Popularity (14)	🟥 Part of Banking Group (8)	🟥 Brand Value (13)
Accessibility	🟥 Two-Factor Authentication (3)			🟩 Two-Factor Authentication (11)		
	🟥 English (9) French (8) Dutch (6) Portuguese (2) Greek (4)			🟥 English (10) French (9) Dutch (3) Polish (4) Hungarian (5) Spanish(2) Swedish (6)		🟩 Italian (7) Portugease (2)

Figure 10.1: Visualisation of the regression results - Red; increasing probability and Green; decreasing probability

.

score as visible in the step-wise model. Only adding this interaction will not change the coefficient from negative to positive since many other interactions are also involved. The assumption is made that bank-size has a positive influence on whether a bank is being targeted.

- **Visibility: Domain popularity, brand value, banking group**
  It appears that visibility of banks in terms of domain popularity, brand value and being part of a banking group are significant characteristics explaining the likelihood to be targeted and its frequency. Especially, *domain popularity* is a very strong predictor whether a bank is targeted and *being part of a banking group* is the best predictor when predicting the target frequency.

- **Two-factor Authentication**
  Having a two-factor authentication has a positive effect on the chance whether a bank will be targeted, but a negative effect on the frequency of it being targeted. Two explanations can be given for this results. First, the descriptive analysis has proved that two-factor authentication correlates largely with certain languages i.e. Swedish and Danish. The collinearity between two-factor authentication and the other independent variables or other underlying factors i.e. cybersecurity maturity could implicate a bias in the estimated coefficients. Secondly, two-factor authentication is not a barrier for adversaries, and thus other criteria are more important to consider for cybercriminals. Yet, some types of the two-factor authentication are indeed effective in reducing targets and it is difficult for criminals to successfully target these banks. Alternatively, banks with two-factor authentication are likely more security mature i.e. they will have better rule-detection systems. This combination causes the actually decrease in targets.

- **Language-use on the website**
  It was expected that largely spoken languages would have an positive effect on target selection, since they ease the attack. However, various - not largely spoken languages - also have an effect on being targeted and on its frequency. The cause of the importance of these languages are mostly related to size, popularity or the level of two-factor adaptation in a country.

4.Which (analysed) banks' characteristics are essential in target selection according to experts?

This question was answered by conducting interviews. The goal of the interviews is to validate the results and gain insight in the practical expertise related to target selection. To validate the results, the experts were asked to what extend they think the results are representative in real world experiences. The results correspond with the practical experiences. Only the results from the two-factor authentication analysis surprised the experts but we agreed on the argumentation.

Other insights, never mentioned in available literature, were gained. These characteristics are: operating at national/international level, the number of provided service provision (investment, business commercial), prominent search results of the bank web-page, allowing on-boarding for digital customers, allowing multiple bank-accounts per citizen. This expands the current list (Table 1.1) structured according to the RAT theory.

- **Value**. Bank-size at a country level: adaption rate of online banking, Financial Status (GDP), ICT development index, rate of banking penetration and the maturity of online banking system.
- **Visibility**. Brand popularity, domain name visibility, banks' attack record, website domain popularity, ownership of the bank, *operating international/ national, prominent search result of the bank web-page and the number of provided services (investment, business commercial).*
- **Accessibility**. Degree of cooperation between financial institutions and law enforcement within a country, the degree of cooperation between financial institutions, authentication method, broadband penetration rate, users' online awareness, security control (in terms of use of firewall/antivirus products and quality of detection system), ease in securely performing criminal actions, language of the online banking and the maturity of the online banking system.
- **Portability**. Ease to recruit money mules, *on-boarding of digital customers, allowing multiple accounts per person,*
  At country level these are: ease of recruiting money mules related to social factors and awareness of customers, money transfer policies of the country.

10.3. The answer to the primary research question

The sub-research questions have been answered and they support the answer to the primary research question.

"Which characteristics of European banks are key in influencing the target selection process of banking malware?"

It has been proved that, next to the occurrence of endogenity, also various banks' characteristics correlate with each other. This affects the estimated coefficients of the regression models. It is therefore difficult to draw valid conclusions based on the output of the regression model. Yet, the results of the regression model in combination with the insights gained by the expert interviews, provide profound knowledge of target selection. This comprehensive understanding enables answering the primary research question.

The first part of the answer contemplates the effect of **bank-size** on target selection. Experts from the banking industry and also an analyst who monitors banking malware, acknowledge the importance of bank-size in target selection. They also envisioned a trend in targeting smaller banks because they have inferior security measures compared to the larger banks. However, proven by the descriptive and regression analyses larger banks have a higher risk of being targeted and are more frequently targeted. Yet, this does not mean that the above statement is false. It needs some nuance and explanation, which will be explained by means of the following findings.

It can be stated that cybercriminals expand their focus of targets. In the years between 2014 and 2017, 20% of the banks signed up for more than 80% of the total number of targets. In the years 2016-2017, 20 % of the banks only attract 70 % of the total number of targets. Furthermore, the descriptive analysis revealed that cybercriminals intended to target specific countries. According to an expert, these criminals first target the larger banks of a country, and when they are successful they also include the smaller banks. After this, they switch to another country. It also has been concluded that cybercriminals remove certain bank names from the configuration file when they are not successful and they add the bank names again after updating their malware. In line in the above identified knowledge, larger banks remain appealing target for cybercriminals, and at some points in time the criminals simultaneously attack the smaller banks to acquire the ”easy money”.

When the bank-size - the value aspect of a suitable target - is compared with the visibility related factors i.e. **domain popularity, brand value and being part of banking group**, the regression analysis proves that these *visibility* factors are better predictors for target selection. The strong prediction of visibility values versus bank-size can be explained as follows. Cybercriminals have bounded rationality, meaning that they have limited information, cognitive limitations, and limited time to make a decision. From bank robbery it is known that criminals select banks from which they hope they would have the highest financial gains. Robbers simply don’t have this kind of information; they make a wild guess which apparently is not always accurate. It is likely that similar patterns are visible in the digital world. Cybercriminals would like to target banks that result in the highest financial gains, but they don’t conduct research on the right information and so they guess which banks are the largest. As a consequence banks with the highest visibility are more vulnerable to be targeted.

The presence of the **Two-factor authentication** has a positive influence on being targeted, but it negatively influences the frequency of being targeted. This controversial relation can be explained. The two-factor authentication is not a barrier for cybercriminals since they can circumvent the security measure. For example, One-Time-Password scheme can be bypassed by using automated MitB attack, i.e. spyrootkits (Adham et al., 2013; Mitnick, 2017). Another method is where the adversary operates in real-time and performs a MitM attack (Windels, 2017). Using those methods, the two-factor authentication is not an obstacle for cybercriminals. The reduction of the lower target frequency can be reasoned by the effectiveness of certain two-factor authentication or complementary security measures that are related with having two-factor authentication. Another reason could be that the estimated coefficient of two-factor is not correct since it highly correlates with various countries and languages. This is expected since the descriptive analyse state that targeted banks have on average less two-factor authentication. Note that this relation is not an causation; it cannot automatically be concluded that two-factor authentication reduces the target frequency.

Experts underlined that the use of the English **language** on the banks’ website eases an cyber attack. Especially in the beginning of the existence of banking malware, banks with largely spoken languages interfaces (i.e. English and German) were more vulnerable: it was not difficult to copy these interfaces. Besides that, these languages are pretty easy to use in combination with soft skills such as social engineering. Currently, interfaces in almost all languages are available. Last but not least adversaries use language services to translate every language they would like to use. This result is in line with the results of the analysis; various not largely spoken languages still influenced the target selection. The cause of the importance of these languages is now more related to the size/visibility or popularity of the banks providing these languages. Moreover, the importance of the English language is likely caused by the international oriented banks since the banking groups was closely related to this language.

10.4. Recommendations

The research intends to support decision-making on technical and policy-based IT security measures by providing a deeper understanding of target selection. This section provides recommendations to perform risk analysis and some recommendations for small and large banks.

It is recommended to incorporate the bank-size metric and the results of this thesis in the banks risk analysis. Most forms of cyber risk analysis require information on the expected loss and the probability of the attack. This research defined that large banks and banks with a high domain popularity are vulnerable to cyber attacks. In the current risk analyses, bank-size is used to assess potential losses. Although the bank-size metric needs some work, this study provides clear insight how the size of a bank can be measured. This metric can be used in a risk-analysis or for the cyber insurance market to determine the insurance premium. The domain popularity is never used in risk assessment, but it is a strong predictor for being targeted. It is recommended to take, next to bank-size, domain popularity into account.

Although smaller banks are not a main target, they are certainly exposed to banking malware attacks. They are in particular vulnerable when they are not keeping up with the latest security protection measures. Smaller banks will probably not have the resources to meet the highest IT security standards. A few options are provided to allocate their scare resources. The first option is to collaborate with other banks or purchase services from third parties, which is already daily practice in Germany and in the UK. The second option is to focus on proper rule-based detection since this is, according to the experts, pretty effective. Even if the two-factor is used as a single security measurement, it is recommended to keep it operational since it does add an extra layer of security (Adham et al., 2013).

Larger banks remain the most attractive targets. Especially for them it is important to innovate on security measures and to keep IT security the cutting edge. For the bank's risk analysis it is recommended to also include the type of services offered to (potential) customers. Allowing customers to digitally open an account, or even worse, allowing them to open multiple online bank accounts may serve the customer satisfaction but the bank is prone to become vulnerable for cybercrime.

Summary - Chapter 10

This chapter answers the primary research question by identifying which characteristics of European banks are key in influencing the target selection process of banking malware. This research proves that large banks, banks that are part of a banking group, banks with a high brand value, and banks which websites have a high domain-popularity, bear a higher probability to be (more frequently) targeted. Two-factor authentication does not decrease the chance for a bank to be targeted. However, the presence of this measure does lead to a lower frequency of attacks. Furthermore, it is shown that banks offering a largely spoken language on their website does not eases the attacks anymore.

In line with the outcomes of this research, some recommendations are proposed for cyber risk analyst, insurance companies, small banks, and large banks. It is recommended to put the bank-size metric and the domain-popularity into practice when performing risk assessments. The bank-size metric and domain popularity (a strong predictor for being targeted) can be used to assess potential losses and attack probabilities for risk analysis or even used to specify (cyber) insurance premiums. For smaller banks, it is recommended to invest in rule-based detection and customer awareness when allocating scarce resources since it experts recognised a huge decrease when implementing rule-based engines. Larger banks should be critical towards allowing digital customer registration or allowing multiple bank accounts for a single customer. These services are necessary for larger banks to increase customers experiences, but they also make the bank prone to banking malware attacks.

The next chapter provides recommendations for future research based on the limitations that are identified in the previous chapter.

11

Future Research

Given the limitations in Chapter 9 and the complexity to create a comprehensive model explaining target selection, an endless number of features could be added and analysed in future research. This section prioritises the possible directions for future research as extension of this research. In line with the limitations described in Chapter 9, this chapter proposes five potential implementations for future research.

Investigate various types of two-factor authentication in relation to target selection

Experts expected that the two-factor authentication would result in a lower chance to be targeted or not targeted at all. It appears that the two-factor authentication is not a saveguard to not be targeted but results in a reduced number of targeted. One explanation is that some two-factor authentications are more effective than others. Future research can differentiate the types of two-factor authentication that are used and research which one results in less targets. Another approach could be to cluster banks in high, medium, and low targeted and identify if those clusters use different types of two-factor authentications. If not; it is recommended to analyse other innovative security measures such as a proper detection to hamper the money laundry process or to look at the general ’security maturity’ of a bank. It would be a challenge to acquire data about a banks’ security maturity or detection transaction monitoring system. This data is not publicly available. A start could be to analyse which banks are listed in the SHIT-list and what these banks have in common in terms of their security measures. Future work is recommended to flesh out these mechanisms more precisely.

Money launder process

Chapter 9 described the importance of ’portability’ related banks’ characteristics in the current model, such as the ease to laundry money. Tajalizadehkhoob et al. (2014) emphasised the need to research money mules or money transfer policies as part of the fraud value chain. Disrupting the money launder process may be a better allocation of scare resources for financial institutions. This suggests research analysts to add portability related factors to the model or investigate the ease of money laundry for different banks/countries. An feasible example is to investigate if enabling the registration of digital customers or allowing multiple bank-accounts per customer influence the target selection process.

Analyse the influence of operating (inter)national to target selection

The regression analysis showed that ”being part of a banking group” positively influences target selection. The underlying general assumption is that these banks operate internationally. Future research could analyse if (inter)national operating banks are more vulnerable for cyber attacks. It is a relatively easy obtainable variable and it is recognised by multiple experts as an important feature of target selection. Operating international/international is therefore recommended for future research.

Extending yearly data

To get a comprehensive understanding of target selection it is important to enhance the model. adding more annual data. The Fox-IT target data is available from 2014-2018, but this research only analysed 2016-2017 to match the financial dataset. Yet, analysing target selection over the period from 2014 till 2018, and thereby using the full potential of the data, will support trend analysis over time.

Analyse mobile malware

Fuelled by the mobile revolution most of the financial transactions are conducted via mobile devices. Two years ago, financial institutions already expected a shift from computer to mobile malware. Since the computer malware is still profitable this shift has not yet been seen. One of the experts argued that it would be helpful to include mobile malware. I would suggest to analyse mobile and computer malware separately. It would prevent that the model becomes over-complicated with to many factors of influence. I would also recommended to perform this analysis for mobile malware in combination with the banks' characteristics from this research as input for the regression model.

Summary - Chapter 11

This chapter proposes five possible directions for future research. The first direction is to investigate the effectiveness of the various types of two-factor authentication. Secondly, add factors to the model that are related to the ease of transferring illegal money, such as: on-boarding of digital customers or the number of accounts per customers. Third, another feature that could be analysed is whether the bank operates (inter)national would increases the vulnerability towards banking malware attacks. Fourth, it is recommended to add more yearly data to the regression model. Ultimately, considering the expected shift from computer to banking malware, it would be recommended to investigate which characteristics influence the target selection in mobile malware.

12 Relevance of this research

This research contributes in several ways to scientific research. It increases understanding in a new research topic of target selection with banking malware. This section starts with explaining the scientific contribution. Not only scientific research benefits from this thesis, but it also adds value to society. Therefore the societal contribution will be explained. The scientific and the societal relevance are composed from the insights obtained in this research. The link with the master programme is explained at the end of the section.

12.1. Scientific contribution

Researches on target selection are a novel research area and not so much of research data is published so far. Most of the studies are fragmented since they focus on different attack vectors, geographic areas or because they are limited to one specific malware. As a consequence, there is no broad and general understanding of the target selection process by criminals who use banking malware. Natalius (2018) made the first step to realise a common understanding in target selection for banking malware by creating an explanatory model including various types of malware.

This study aims to develop a more complete explanatory model towards target selection in banking malware and, consequently, increases the quality of the model and research. This research was able to contribute to the scientific literature in four ways.

First, it introduced a bank-size metric for target selection studies. This selection is done by a thorough analysis of the advantages and disadvantages of each of the eleven measurements. This information and outcomes can easily be used for other (cyber-) studies that would like to use the bank-size metric. In addition, the five most suitable measures for ”being targeted” and ”target frequency” are identified. The method used to identify the most suitable bank-size measures is explained and can be reproduced if desired.

Second, the research identified the influence of bank-size on target selection. Financial and market factors are fundamental explanations of target selection in the banking malware. In this way, the explanatory model was able to contribute to the RAT by exploring if ’value’ is an element that can explain victimisation by banking malware.

Third, this research took it even a step further by also exploring ’being part of a banking group’ and ’brand-value’. For future research, it is clear that those factors should be analysed.

Fourth, this research makes all data and python code publicly available to stimulate future target selection studies. Researchers will probably use new or different methods of analysis. With this help, they can do so without getting too much hassle on the availability of the data.

This research benefited from data generated in previous research work, and I hope others can benefit from my collected data.

12.2. Societal relevance

This research identified which bank characteristics make a bank vulnerable to cyber-related attacks. It showed trends in banking malware and visualised in figure 10.1 which characteristics decrease/increases the banks' cyber security risk. This knowledge supports banks in formulating a proactive security policy and supports decision-making about the technical and policy-based risk mitigation strategies. Multiple experts in the field acknowledged the importance of such a model. They also confirmed that the explanatory model could support a better focus on security measures and could even support threat intelligence operations. Next, this research recommends in Section 10.4 which actions could decrease the effects of banking malware.

Most cyber security reports use the Raw count metric, but this metric has limitations. The researched Unique count metric seemed to be more reliable. If cyber security companies would use this metric, their reports and analyses may provide deeper insights. The metric of Unique count is defined in earlier research. However, these studies didn't make a distinction between malware threats.

12.3. Link to master programme

The Engineering and Policy Analysis (EPA) master programme aims to use analytical and modelling techniques to support decision-making.Many of these techniques are used throughout the study to get a better and deeper understanding of the problem and to work towards scientifically supported conclusions. The efforts can be confirmed by the number of python and r packages, Appendix C, that have been used to complete the data preparation and regression model. This research uses analytical and modelling tools to provide intelligence on which banks characteristics are attractive for cybercriminals who deploy banking malware. This intelligence supports decision-making for the financial sector to prioritise and allocate their resources to prevent banking malware attacks. Moreover, this intelligence is useful for cyber risk analysts to asses risk more effectively and for insurance companies to determine insurances rates.

In addition, one of the core contents of the EPA programme is to analyse grand challenges of todays' world. This research uses analytical and modelling tools to understand target selection in banking malware which contributes to understanding the grand challenge of cybersecurity.

Summary - Chapter 12

This chapter identifies how this research is relevant to society and scientific research. Starting with explaining the scientific contribution, this study is the first research to introduce a bank-size metric and explores the relation of bank-size towards target selection. Moreover, it added new variables to the regression model: brand value and being part of a banking group. In addition, this research makes the financial dataset, python and R code, publicly available to encourage future research.

This research contributes to society by defining an adequate metric to measure target frequency for cybersecurity companies. Furthermore, the model is able to support threat intelligence operations and risk assessments.

This research relates to the master programme by exploring the field of cybersecurity, which is defined as one of the grand challenges of today's world. It uses analytical and modelling tools to support decision-making for various experts in the field of (cyber) risk assessment, insurance, and banking.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

Acurna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect in the classifier accuracy, classification, clustering, and data mining applications. In *Proceedings of the meeting of the international federation of classification societies (ifcs)* (pp. 639–647).

Adham, M., Azodi, A., Desmedt, Y., & Karaolis, I. (2013). How to attack two-factor authentication internet banking. In *International conference on financial cryptography and data security* (pp. 322–328).

Allisy-Roberts, P., Ambrosi, P., Bartlett, D. T., Coursey, B. M., DeWerd, L., Fantuzzi, E., & McDonald, J. (2018). *2018 Internet Security Threat Report* (Vol. 6; Tech. Rep. No. 2). Retrieved from https://academic.oup.com/jicru/article-lookup/doi/10.1093/jicru/ndl025 doi: 10.1093/jicru/ndl025

Arbor Networks. (2016). *Worldwide Infrastructure Security Report 2016 (Vol. XII)* (Tech. Rep.).

Arinaminpathy, N., Kapadia, S., & May, R. M. (2012). Size and complexity in model financial systems. *Proceedings of the National Academy of Sciences*, *109*(45), 18338–18343. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.1213767109 doi: 10.1073/pnas.1213767109

Asociación Española De Banca. (2017). Anuario estadístico de la Banca en España.

Bankier.pl. (2017). Polska bankowość w liczbach i kw.2017.

Barnum, S. (2014). Standardizing cyber threat intelligence information with the Structured Threat Information eXpression (STIX™). *MITRE Corporation, July*, 1–20. doi: 10.1002/ejoc.201200111

Baron, P. (2015). The 10 Most Important Blood Tests. *Life Extension Magazine*(May). Retrieved from http://www.lifeextension.com/Magazine/2006/5/report{_}blood/Page-01

Beazley Breach. (2016). Beazley Breach Insights. (July), 1–2.

Black, P., Gondal, I., & Layton, R. (2017). A survey of similarities in banking malware behaviours. *Computers & Security*. Retrieved from https://doi.org/10.1016/j.cose.2017.09.013 doi: 10.1016/j.cose.2017.09.013

Box, G. E. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, *143*(4), 383–404.

Brandirectory. (n.d.). *Banking 500 2017.* Retrieved 2019-03-05, from https://brandirectory.com/rankings/banking-500-2017

Bulgarian National Bank. (2017). Banks in Bulgaria. (June).

Cangelosi, R., & Goriely, A. (2007). Component retention in principal component analysis with application to cdna microarray data. *Biology direct*, *2*(1), 2.

Carter, W. A. (2017). FORCES SHAPING THE CYBER THREAT LANDSCAPE FOR FINANCIAL INSTITUTIONS. (November), 1–31.

Centre, N. C. S. (2017). *Cyber crime: understanding the online business model* (Tech. Rep.). Author.

Charalambous, G. (2018). *Development of Injected Code Attacks in Online Banking Fraud Incidents* (Unpublished doctoral dissertation).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chebyshev, V., Sinitsyn, F., Parinov, D., Liskin, A., & Kupreev, O. (2018). *IT threat evolution Q1 2018. Statistics.* Retrieved 13-08-2018, from https://securelist.com/it-threat-evolution-q1-2018-statistics/85541/

Chen, D. X., Office, S., Damar, H. E., & Soubra, H. (2012). An Analysis of Indicators of Balance-Sheet Risks at Canadian Financial Institutions. (2011), 21–33.

Cheung, R. (2017). *Targeting financial organisations with DDOS: A multi-sides perspective* (Doctoral dissertation, Delft University of Techology). Retrieved from https://repository.tudelft.nl/islandora/object/uuid{%}3Acd5fd5ea-ec4e-4cf4-bb8c-2fc4c84f583a

Creswell, J. W., & Clark, V. L. P. (2011). *Designing and conducting mixed methods research.* SAGE.

Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, *28*(1), 224–238.

Cucu, P. (2017). *Rootkit – the (Nearly) Undetectable Malware.* Retrieved 2-10-2018, from https://heimdalsecurity.com/blog/rootkit/

Custers, B. H., Pool, R. L., & Cornelisse, R. (2018). Banking malware and the laundering of its profits. *European Journal of Criminology*, 147737081878800. Retrieved from http://journals.sagepub.com/doi/10.1177/1477370818788007 doi: 10.1177/1477370818788007

Dang, C. D., & Li, F. (2015). Measuring Firm Size in Empirical Corporate Finance Measuring Firm Size in Empirical Corporate Financ e Abstract. (519).

de Bruijne, M., van Eeten, M., Gañán, C. H., & Pieters, W. (2017). Towards a new cyber threat actor typology. *Delft University of Technology*.

de González, A. B., Cox, D. R., et al. (2007). Interpretation of interaction: A review. *The Annals of Applied Statistics*, *1*(2), 371–385.

de Groen, W. P. (2016). *" Total Assets " versus " Risk Weighted Assets ": does it matter for MREL requirements ?* (Tech. Rep.). European Parliament.

Deutsche Bundesbank Eurosystem. (2017). *Banken und andere finanzielle Institute.* Retrieved 2019-03-05, from https://www.bundesbank.de/de/statistiken/banken-und-andere-finanzielle-institute

Die bank. (2017). TOP 100 DER DEUTSCHEN KREDITWIRTSCHAFT Zartes Erwachen Euro. *Die Bank*(8), 12–19. Retrieved from https://www.wiso-net.de/document/DIBA{_}{_}2017080169

Dixon, J. K. (1979). Pattern recognition with partly missing data. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(10), 617–621.

Duncan, B. (2018). *Malware Team Up: Malspam Pushing Emotet + Trickbot.* Retrieved 2019-03-05, from https://unit42.paloaltonetworks.com/unit42-malware-team-malspam-pushing-emotet-trickbot/

European Central Bank. (2018). *List of Monetary Financial Institutions.* Retrieved from https://www.ecb.europa.eu/stats/financial{_}corporations/list{_}of{_}financial{_}institutions/html/index.en.html

Europol. (2017). Banking Trojans : From Stone Age to Space Era. , 1–16.

Eurostat. (2018). *Individuals - internet activities.* Retrieved 17-12-2018, from http://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK{_}DS-053730{_}QID{_}4CBC0374{_}UID{_}-3F171EB0{&}layout=TIME,C,X,0;GEO,L,Y,0;INDIC{_}IS,L,Z,0;UNIT,L,

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences.* Brooks/Cole.

Febelfin. (2017). *The Data Driven Bank.* Retrieved from https://www.febelfin.be/en/data-driven-bank

Financial Fraud Action UK. (2018). *Fraud the Facts 2018: The definitive overview of payment industry fraud* (Tech. Rep.).

Florêncio, D., & Herley, C. (2010). Phishing and money mules. In *2010 ieee international workshop on information forensics and security* (pp. 1–5).

Gardiner, J., Cova, M., & Nagaraja, S. (2014). Command & Control: Understanding, Denying and Detecting. *arXiv.org*, *cs.CR*(February), 38. Retrieved from http://arxiv.org/abs/1408.1136v1{%}5Cnpapers3://publication/uuid/4389EDB2-DA22-4672-8B1F-A0F60556CA73

German Savings Bank Association. (2016). Rangliste 2016. (May), 31–48.

Hilbe, J. M. (2011). *Negative binomial regression.* Cambridge University Press.

Hrvatska Narodna Banks. (2018). *Credit institutions.* Retrieved from https://www.hnb.hr/core-functions/supervision/list-of-credit-institutions/-/

asset{_}publisher/e66e1a1e3a/content/credit-institutions

IBM Software. (2014). Financial malware explained. (December).

Ilvento, T. (n.d.). Using Statistical Data to Make Decisions Using Statistical Data to Make Decisions: Multiple Regression Analysis Module 5: Multiple Regression Analysis BASICS OF MULTIPLE REGRESSION.

Jaishankar, K. (2011). *Cyber Criminology*. Retrieved from http://www.crcnetbase.com/doi/book/10.1201/b10718 doi: 10.1201/b10718

Jones, J. A. (2005). *An Introduction to Factor Analysis of Information Risk (FAIR)* (Tech. Rep.). Risk Management Insight.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895–2907.

Khan, S. R. (2018). Implication of Cyber Warfare on the Financial Sector. An Exploratory Study. *International Journal of Cyber-Security and Digital Forensics*, *7*(1), 31–38. Retrieved from http://go.galegroup.com.proxy.library.cmu.edu/ps/i.do?{&}id=GALE{%}7CA535538243{&}v=2.1{&}u=cmu{_}main{&}it=r{&}p=AONE{&}sw=w doi: 10.17781/P002319

Kiljan, S., Vranken, H., & van Eekelen, M. (2018). Evaluation of transaction authentication methods for online banking. *Future Generation Computer Systems*, *80*, 430–447.

Klason, A. (2018). *Bokbot: The (re)birth of a banker.* Retrieved 2019-03-05, from https://blog.fox-it.com/2018/08/09/bokbot-the-rebirth-of-a-banker/

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).

Laeven, L., & Ratnovski, L. (2014). Bank Size , Capital Requirements , and Systemic Risk : Some International Evidence.

Laeven, L., Ratnovski, L., & Tong, H. . (2014). Bank size and systemic risk: Some international evidence. *International Monetary Fund, Mimeo.*, *69*.

Le principali banche Italiane. (2016). Le principali banche Italiane. (6). Retrieved from http://www.mbres.it/sites/default/files/resources/download{_}it/ps{_}6{_}8.pdf

Lepetit, L., Nys, E., Rous, P., & Tarazi, A. (2008). Bank income structure and risk: An empirical analysis of european banks. *Journal of Banking & Finance*, *32*(8), 1452–1467.

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

Lund, M. S., Solhaug, B., & Stølen, K. (2010). *Model-driven risk analysis: the CORAS approach.* Springer Science & Business Media.

Marques-ibanez, D. (2012). Bank Ratings What Determines Their Quality ? (1484).

McFadden, D., et al. (1977). *Quantitative methods for analyzing travel behavior of individuals: some recent developments.* Institute of Transportation Studies, University of California Berkeley.

McKillop, D., O'Connell, D., & O'Toole, J. (2016). A Survey of Irish Credit Unions. (January), 1–24.

Mikhaylov, A., & Frank, R. (2017). Cards, money and two hacking forums: An analysis of online money laundering schemes. In *Proceedings - 2016 european intelligence and security informatics conference, eisic 2016.* doi: 10.1109/EISIC.2016.021

Miles, C., Lakhotia, A., LeDoux, C., Newsom, A., & Notani, V. (2014). Virusbattle: State-of-the-art malware analysis for better cyber threat intelligence. In *Resilient control systems (isrcs), 2014 7th international symposium on* (pp. 1–6).

Mitnick, K. (2017). *Evilginx - Advanced Phishing with Two-factor Authentication Bypass.* Retrieved 2019-03-05, from https://breakdev.org/evilginx-advanced-phishing-with-two-factor-authentication-bypass/

Morrison, S. (1996). An Analysis of the Decision Making Processes of Armed Robbers. *The Politics and Practice of Situational Crime Prevention*, *5*, 159–187.

Nagunwa, T. (2014). Behind Identity Theft and Fraud in Cyberspace : The Current Landscape of Phishing Vectors. , *3*(1), 72–83.

Natalius, S. (2018). Assessing the Role of Online Banking's Characteristics in the Target Selection of the Banking Malware. (August).

OECD Committee on Financial Markets. (2017). Lithuania : Review of the Insurance System. (November).

Pratt, T. C., Holtfreter, K., & Reisig, M. D. (2010). Routine online activity and internet fraud targeting: Extending the generality of routine activity theory. *Journal of Research in Crime and Delinquency*, *47*(3), 267–296. doi: 10.1177/0022427810365903

PWC. (2011). Banking Profitability and Performance Management. *Banking Profitability and Performance Management*, 1–17.

Ringle, C. M., Wende, S., & Becker, J.-M. (2015). *Smartpls 3. boenningstedt: Smartpls gmbh.*

Romanosky, S., Ablon, L., Kuehn, A., & Jones, T. (2017). Content Analysis of Cyber Insurance Policies: How Do Carriers Write Policies and Price Cyber Risk? *Ssrn*, 1–38. doi: 10.2139/ssrn.2929137

Rosenquist, M. (2009). Prioritizing Information security risk with threat agent risk assessment. *Intel White Paper*, 8.

Saleh, M. F. (2011). Information security maturity model. *International Journal of Computer Science and Security (IJCSS)*, *5*(3), 21.

Schildbach, J., & Schneider, S. (2017). *Large or small? how to measure bank size.*

Shackleford, D. (2015). Who's using cyberthreat intelligence and how? *SANS Institute. Retrieved February*, *23*, 2016.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161–176.

Šimundić, A.-M. (2009). Measures of diagnostic accuracy: basic definitions. *Ejifcc*, *19*(4), 203.

Singer, J. D., Willett, J. B., Willett, J. B., et al. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford university press.

Ståhlberg, M., & Corporation, F.-s. (2007). THE TROJAN MONEY SPINNER STÅHLBERG THE TROJAN MONEY SPINNER. (September), 1–7.

Swedisch Bankers. (2017). *Commercial banks, December 31, 2017* (Tech. Rep.).

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Symantec. (2015). Attackers Target Both Large and Small Businesses. , 40.

Tajalizadehkhoob, S. (2013). Online Banking Fraud Mitigation. *Business Week*.

Tajalizadehkhoob, S., Asghari, H., Gañán, C., & Eeten, M. V. (2014). Why Them ? Extracting Intelligence about Target Selection from Zeus Financial Malware 1 Introduction. , 1–26.

Tajalizadehkhoob, S., Böhme, R., Ganán, C., Korczyński, M., & Eeten, M. V. (2018). Rotten apples or bad harvest? what we are measuring when we are measuring abuse. *ACM Transactions on Internet Technology (TOIT)*, *18*(4), 49.

The Banker Database. (2018). *Sample Data.* Retrieved from https://www.thebankerdatabase.com/

The MITRE Corporation. (2017). *Target Selection - pre-attack.* Retrieved 2018-08-11, from https://attack.mitre.org/pre-attack/index.php/Target{_}Selection

ThreatConnect. (2016). *Gootkit Banking Malware.* Retrieved 2019-03-05, from https://threatconnect.com/gootkit-banking-malware/

Van Moorsel, D. (2016). *Target selection regarding financial malware attacks within the Single Euro Payments Area* (Doctoral dissertation, Delft University of Technology). Retrieved from http://repository.tudelft.nl/islandora/object/uuid:c1c0e0e0-7fe2-469d-a010-325f1942f89b/datastream/OBJ/download

Waemustafa, W., & Sukri, S. (2016). Systematic and unsystematic risk determinants of liquidity risk between Islamic and conventional banks. *International Journal of Economics and Financial Issues*, *6*(4), 1321–1327. doi: 10.15408/aiq.v8i2.2871

Weisel, D. L. (2007). *Bank Robbery* (No. 48).

Windels, J. (2017). *How hackers are using phishing to bypass two-factor authentication.* Retrieved 2018-06-29, from https://www.wandera.com/mobile-security/bypassing-2fa/

Wright, R., & Decker, S. H. (2002). Robbers on Robbery: Prevention and the Offender.

Wueest, C. (2017). *Financial Threats Review 2017* (Tech. Rep.). Symantec. Retrieved from https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/istr-financial-threats-review-2017-en.pdf

Yar, M. (2005). The novelty of ‘cybercrime’ an assessment in light of routine activity theory. *European Journal of Criminology*, *2*(4), 407–427.

A

Banking malware families

In general, malware can be distinguished in terms of their business models: **Private, Rented and Kit malware.** The Kit malware, such as Zeus - including a basic support by the seller of the malware - can be bought on the underground forums for roughly 10.000 dollars. Rented malware is often written by technically skilled criminals. They rent out their services on underground forums. The buyer of the Rented malware pays a montly fee to operate the malware. Private malware is mostly developed and operated by a professional and highly-skilled individual or criminal group for their own purposes. Private malware can also use elements of Kit malware.

Kit malware

The following malware is distinguished as Kit malware, or in other words, a malware-as-a-service variant.

- **Zeus** or Zbot was first detected in 2006 and is the first attack vector that applied MitB techniques (Black et al., 2017). The malware runs on Microsoft Windows and is often deployed to steal banking information by MitB, keystroke logging and form grabbing. Zeus is also used as a distribution malware to install the CryptoLocker ransom-ware. It is usually spread by drive-by downloads and phishing schemes. A variety of malware is based on the Zeus code and then belongs to the Zeus malware family. The following variants of Zeus can be defined and will be explained below: Citadel, Zeus Panda, KINS and Ice-IX.

  - **Citadel** is based on the leaked code of Zeus source code and detected in 2011. It operates as a malware-as-a-services business and can be bought on Russian underground forums. Over the years, it compromised millions of computers around the world Europol (2017), but mostly targeted financial institutions located in the Europe, UK, Germany and the Netherlands (Tajalizadehkhoob, 2013). The payload of the malware is modified over time to deal with a large variety of websites and services, including 'Password safe' or 'Keepas'. Also, it targeted webmail services such as Gmail, Yahoo, Japan mail and hotmail.
  - **Zeus Panda** is also a variant of the Zeus trojan horse. It is discovered in 2016 and carries the same coding as Zeus to execute MitB, keystroke loggin, and form grabbing. It also uses the same methods of infection, namely drive downloads, poisoned email and word document macros. It uses other malware, such as Emotet, Smoke Loader, Godzilla dan Hancitor, to distribute Zeus Panda into the system. Zeus Panda is able to target systems in specific regions of the world. It detects and counters many forensic analytic tools and sandbox environments.

    Panda is often used for financial services in Europe, but before the Olympic Games it ran a special campaign towards the Brazilian banks.

  - **KINS** is malware, short for Kasper Internet Non Security, is fully based on the ZeuS code with some minor adaptations. KINS targets financial institutions in Europe, in particular in Germany and The Netherlands. The additional functionality is that it is able to report installed security product information. Also, it added software/code aimed to complicate malware analysis.
  - **ICE-9/Ice IX** is a botnet that uses the original ZeuS code to steal users banks account details and even phone numbers. The infected file may come from a multitude of sources including: floppy diskettes, downloads through an online service, network, etc.

- **Matrix** The initial loader for Matrix Banker sets persistence through Registry Run, and extracts and injects a DLL into Chrome, Firefox, Internet Explorer or Edge. Once the main DLL is injected a browser, it starts by hooking the appropriate browser functions (e.g. PR_Read and PR_Write for Firefox) to setup a MitB. It then phones home to its C2 server to get the webinject config. The request looks like this: Bleeping Computer, a victim needs only to visit a website running the malicious advertising while running the unpatched software to become infected

- **ReactorBot** Corebot or reactorbot, was mainly active in the summer of 2015 and has been spread once again with an updated version. The infection vector of the reactorbot is a spam e-mail with malicious office documents as attachments. Then, the malware uses form grabbin, which contains malicious JavaScript, that is able to harvest details from the victim. Corebot used a common web-inject framework, that are similar to Tinba and Gozi families.

- **Kronos** Kronos is a Banking Trojan and geo-targeted attacks to Australia, Italy, UK and US. It started by sending money mule spam to users with the country-code domain.

- **Ramnit** malware is a worm that first appeared in 2010. In 2011, due to the leak of ZeuS source code, Ramnit obtained some of ZeuS' functionalities and performs MitB attacks. The difference compared to the ZeuS web injects module, is that Ramnit does not communicate with the C&C server directly. Throughout the years, the Ramnit distribution is growth and is even in the top list of malware (bron). In 2015, a take-down operation is conducted, but failed and it is still a running campaign, mostly focused on the customer of major banks in the UK.

- **PkyBot** is a Banking Trojan that distributes Bublik in 2013 and GameOverZeus in 2014. In 2015, also web inject capabilities where added.

- **NuclearBot** or TinyNuke is a Banking Trojan which main functionality is to make web injections into specific pages to steal user data. The source code is available on Github.

- **Gozi, Gozi-EQ, Gozi-ISFB**. In 2007, the Gozi malware is discovered and targeted mostly English speaking countries. It is just like Zeus, a malware-as-a-service Trojan that infects victims using a rootkit. The source code is leaked in 2010, so the author tried to add web injection capabilities (Europol, 2017). The second variant of the Gozi malware was able to inject code when targeting different banks. The developers of Gozi malware have updated the code of web injection mechanism to commit ”form grabbing” and web injection in the Edge browser which is the default browser of Windows 10(Charalambous, 2018). The Gozi-EQ malware used multiple steps to perform an attack. This also accounts for ZeuS-P2P, and differ from Citadel and Ramnit.

- **Tiny Banker Trojan (Tinba-v1, Tinba-v2)** is a highly modified version of the Zeus banker trojan and contains similar attack methods to obtain information. However, Tinba is smaller and more powerful, the rootkit capabilities and size makes it difficult to detect. The code is very simple and it does not contain any advanced encryption methods Europol (2017). It uses MitB attack and network sniffing. Tinba grabs keystrokes through form grabbing therefor they can be encrypted by HTTPS and sends these keystrokes to the C2. Popular distribution methods are malvertising, spam campaigns, and exploit kits. Some of the campaigns are geo-specific, but victims are all around the world.

Private Malware

- **Retefe-v2** is a Windows Banking Trojan that can download and install additional malware onto the system using Windows PowerShell. The goal is to assist the adversary to steal credentials for online banking websites. It is typically targeted Austria, Sweden, Switzerland, UK and Japan. It operates by routing traffic involving the targeted banks through its proxy. Retefe is delivered via zipped .js (JavaScript) files and, as in this particular campaign, via Microsoft Word documents.
- **Nymaim** is a malware dropper and credential stealer. Nymaim was discovered in 2013 and mainly targeted Poland, Germany and the United States. The main function is to funnel additional malware (mostly ransom-ware) onto the infected system trough drive-by downloads, malicious links and email attachments.
- **Qakbot** is a worm (worm replicate functional copies of themselves and are stand-alone software) that spreads through network shares and removable drives. It downloads additional files, steals information, and opens a back door on the compromised computer. The worm also contains rootkit functionality to allow it to hide its presence. The Qakbot spreads by exploiting vulnerabilities when a user visits certain Web pages. The worm also spreads through network shares by copying itself to shared folders when instructed to by a remote attacker. It also copies itself to removable drives. Qakbot is capable of gathering Authentication cookies, including Flash cookies, DNS, IP, host name details OS and system information, Geographic and browser version information Keystrokes including login information, Login details for FTP, IRC, POP3 email, and IMAP email, Outlook account information, Private keys from system certificates, Login credentials for certain websites, URLs visit.
- **Dyre**

  Dyre malware is detected in 2014 and is known by the following names: Dyreza, Dyzap or Dyranges. It targets mostly English language speaking countries. The Dyre malware is able to execute different types of MitB attacks. The first step is similar to Zeus, where malicious code is injected in order to redirect victims to a fake website (on Chrome, Firefox or Internet Explorer) and injects his credentials. Throughout 2014, Dyre used the Upatre downloader to infect its victims' computers. After the info-stealing phase, Dyre frequently deploys other malware on the victim machines. In many cases, it turns them into spam bots which can be used for future infections and various other malicious activities.Compared to other banking Trojan malware, the Dyre malware has no connection with the ZeuS malware family, but some claim that it is connected to Gozi (Europol, 2017).

  Dyre has the full package of banking Trojan capabilities. An additional feature enables it to bypass a website's SSL mechanism by redirecting the victim's traffic as uncrypted while still displaying on an HTTPS encrypted session mark on the victim's web browser. This deception technique is highly effective against security-conscious victims Europol (2017). Furthermore, it has a Invisible Internet Project (I2P) tunnelling support Charalambous (2018). This is an anonymous overlay network, thus a network in a network and intended to protect communication. Last, it has a Built-in Virtual Network Computing (VNC) server. This is a graphical desktop sharing system that uses the Remote Frame Buffer (RFB) protocol in order to control another computer remotely.
- **Rovnic** malware is distributed via macro downloader, which is previously seen in the Dridex malware. The malware target for 95 per cent Germany.

Rented malware

- **Qadars** is a trojan used for banking fraud and credential theft. It targets users via exploit kits and install Powershell Script (programming language). It is able to steal infected users two factor authentication codes and banking credentials through the deployment of webinjects.

  capabilities

  - The C2s are not utilised solely for the collection of stolen credentials. We have also observed them delivering a module to Qadars samples operating in a low privilege environment that employs social engineering to trick the user into allowing higher level access.
  - DGA

- **Dridex** Dridex bot is one of the most adaptable and prevalent Banking Trojans that is active from 2014. Dridex is based on the Cridex, a worm that spread itself through network drivers and external devices. Dridex botnet is divided into subnets that all have different delegations. The source code of the bot is continuity updated by the developers . Dridex exploit the vulnerabilities of the Chrome web-browser. Dridex look up chrome version in the associate uninstall registry key.

  Dridex focused on English-speaking countries and had a pretty high infection rate, compared to other Banking Trojans Europol (2017). The success of the malware lies in its distribution method. The malware is spread by spam campaigns which run at a high rate. The messages include malicious macros (Microsoft or excel), like kronos and rovnic, which download the Dridex. Then, bank information for the software installs a keyboard listener and performs injection attacks. Often victims that gets financial information through email (excel attachments) were a victim. In 2015, thousands machines were added to the Dridex botnet on a daily basis.

- **ZeuS-P2P** Zeus-P2P or GameOVer Zeus appeared in 2011 and is obviously a variant of Zeus. The evolutionary attribute of this type is the communication method, namely peer-to-peer Europol (2017). This means that the interconnected nodes (peers) share resources among each other without the use of centralised system. The bots communicate every 30 minutes with its neighbouring bots and tries to contact other bots when there is no response. In October 2014, the network contained at least 200,000 bots. Since the malware does not require a centralised C2, it is very difficult to detect its activity. On the other hand, it still uses DGA as a back up communication method.

B

Sources Financial Database

This appendix shows the utilised sources on country-level. The sources per European bank can be found in the column ”Source” in the financial database. Here, also the year of the sources is described. This database is made publicly availibly on: https://github.com/MarritH/TargetSelection_BankingMalware.

Table B.1: Overview of financial database sources on country-level

Country	Source
All	The Banker Database (2018)
Belgium	Febelfin (2017)
Bulgaria	Bulgarian National Bank (2017)
Croatia	Hrvatska Narodna Banks (2018)
Germany	Deutsche Bundesbank Eurosystem (2017); Die bank (2017); German Savings Bank Association (2016)
Italy	Le principali banche Italiane (2016)
Poland	Bankier.pl (2017)
Spain	Asociación Española De Banca (2017)
Sweden	Swedisch Bankers (2017)

C

Utilised Python and R Packages

In Chapter 5 and 6 the programming language Python is used to prepare and visualise the data. For this, the scientific python development environment *spyder* is used. In chapter 7 the programming language R is used to build and analyse the regression models. All python and R-code can be found on the github page:
`https://github.com/MarritH/TargetSelection_BankingMalware.`

Table C.1: Utilised Python packages and the version numbers

Name	**Version**	**Usage**	**Chapter/Section**
Imbalanced-learn	0.3.1.	Dealing with an imbalanced dataset using SMOTE	Section 7.3
FancyImpute	0.4.2	Data Imputation and Elbow Methode	Section 5.3.1
Folium	0.7.0	Mapping the geographical distribution	Figure 6.1, 6.2 and 6.3
FuzzyMuzzy	0.17.0	Combine Datasets on Bank Name	Section 5.4
Matplotlib	2.1.0	Creating figures	Chapter 6
Notebook	5.0.0	Make visualisation	Chapter 6
Numpy	1.14.2	Array processing for numbers, strings, records and objects	N.A.
Pandas	0.20.3	Structure and analyse the data	N.A.
Spyder	3.2.4	Scientific python development environment used to clean and prepare the data	
Scikit-learn	0.19.1	Make a stratified sample, normalise the Data	Section 5.2
Scipy	1.0.0	Interpolate the Data	Section 5.3.1
Seaborn	0.9.0	Statistical data visualisation	Chapter 6
Statsmodels	0.8.0	Variance Inflation factors, Interpolation	Chapter 7
System		Python 3.6.0 |Anaconda custom (64-bit)| [MSC v.1900 64 bit (AMD64)]	

Table C.2: Utilised R packages for Regression Analysis

Name	**Chapter/Section**
Ggplot2	Ggplot is a data visualization package for the statistical programming language R
Factoextra	Extract and Visualise the Results of Multivariate Data Analyses
MMASS	Modern Applied Statistics with S'
Prin_comp	Build-in R programme to build Principal Component Analyse
Boot	Calculate equi-tailed two-sided nonparametric approximate bootstrap confidence intervals for a parameter, given a set of data and an estimator of the parameter, using numerical differentiation.

D

Visual graphs: Interpolation

Figure D.1, D.2, D.3, D.4 and D.5 show the data-points of the financial database (dark-blue) and the data-points that are created with the interpolation method (light blue).

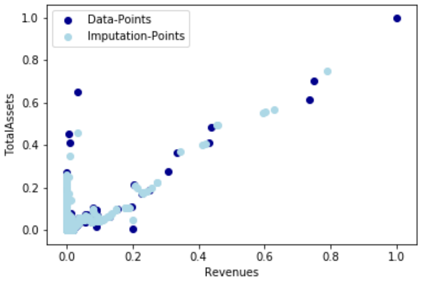


Figure D.1: Interpolation points Revenues

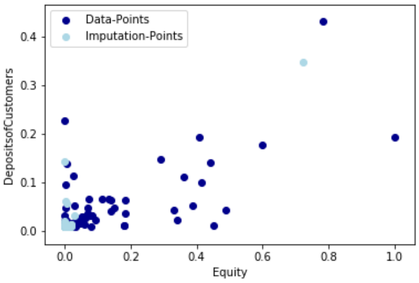


Figure D.2: Interpolation points Equity

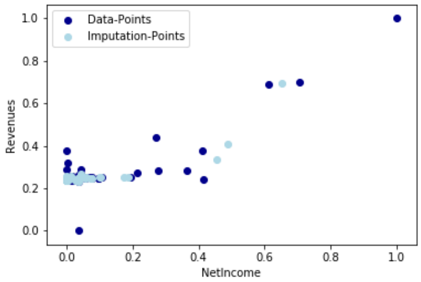


Figure D.3: Interpolation points Net Income

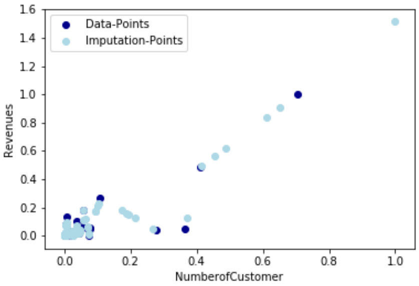


Figure D.4: Interpolation points Number of Customers

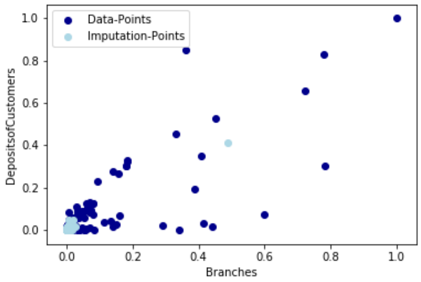


Figure D.5: Interpolation points Branches

E

Visual graphs: Descriptive analysis

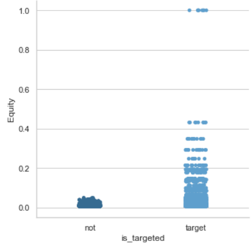


Figure E.1: Equity

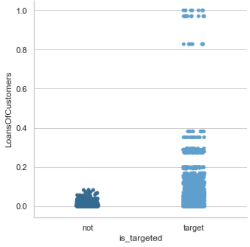


Figure E.2: Loans of Customers

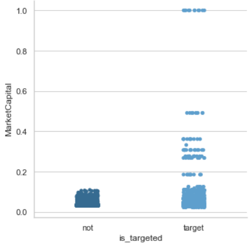


Figure E.3: Market Capital

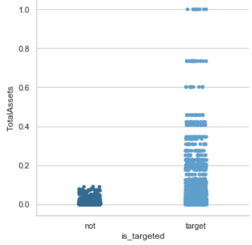


Figure E.4: Total Assets

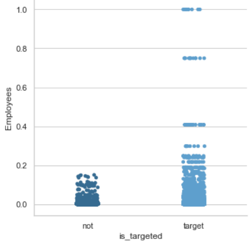


Figure E.5: Number of Employees

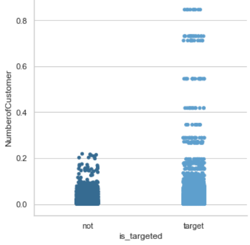


Figure E.6: Number of Customers

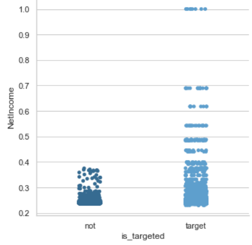


Figure E.7: Net Income

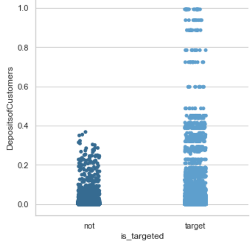


Figure E.8: Deposits of Customers

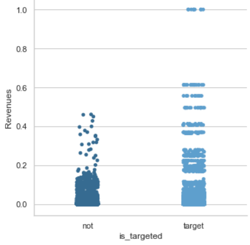


Figure E.9: Revenue

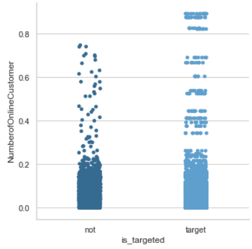


Figure E.10: Number of Online Customers

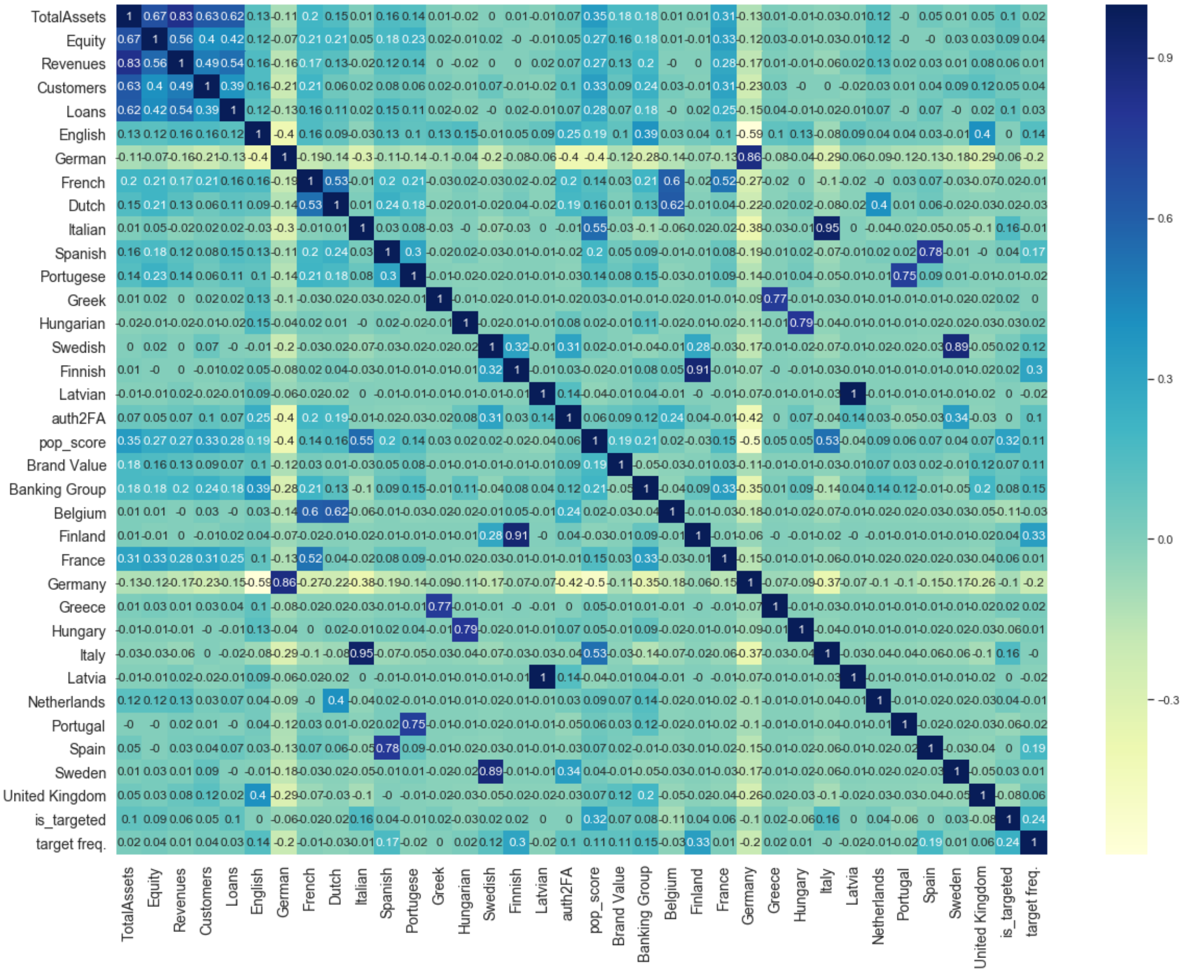


Figure E.11: Correlation Matrix; banks' characteristics

F

Feature Selection

It appears that the financial data suffers from multicollinearity. Multicollinearity can be measured using VIF. This appendix identifies and removes irrelevant and redundant information on the bank-size measures by performing feature selection. The goal is to exclude less important, irrelevant or distracting variables for measuring bank-size resulting in more constant results and less collinearity in the dataset.

F.1. Feature Selection methods

Similar to the regression analysis, the Feature Selection analysis distinguishes two dependent variables, namely ”being targeted” (dichotomous) and ”target frequency” (continuous). Different techniques will therefore be applied. The bank-size measures are the independent continuous variables.

A variety of methods can be used to reduce the number of bank-size measures using feature selection. One method is to choose the model with the highest relative measure of fit: R-squared. However, this would not be a reliable approach for the selection of the best model since the R-squared value will always increase when more features are added. Therefore I'll choose to apply feature/variable selection techniques: **filters** and **wrappers**. Filters select a subset of variables independent of the model that will later use them. The results will exclude some non-predictable bank-size measures in order to increase the efficiency of the wrapper methods. The wrappers select a subset of variables in the model. The embedded method is built in the ML model (or rather its training algorithm). A complete and comprehensive description can be found below:

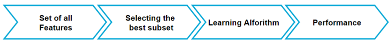


Figure F.1: Filter

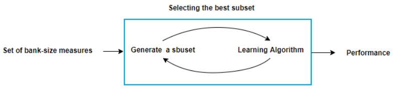


Figure F.2: Wrapper method

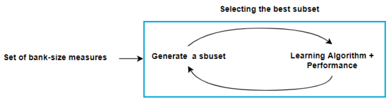

Selecting the best subset
Set of bank-size measures
Generate a sbuset
Learning Algorithm + Performance

Figure F.3: Embedded method

F.1.1. Filters

The **filter techniques** are mostly based on selecting variables through the use of between-class separability criteria. These techniques do not consider the effect of selected variables on the performance of an entire processing algorithm (see figure F.1). It does not involve predictive evaluation for reduced data-sets with selected variable subsets only. Still, the filter techniques are useful to be the first selection step before performing the wrapper analysis. The wrapper techniques are computational intensive and a selection beforehand can increase the performance. Wrapper technique uses predictor performance as a criterion of variable subset selection. It is expected that the wrapper provides a better subset selection, since it identifies the optimal variables selection.

Filter methods provide a set of 'the most important' variables independent of the employed model.
The Filter methods distinguish two types of variables; continues and categorical. As mentioned, this research explores the relation of the independent variables and the dependent variables; target frequency (continuous) and being targeted (categorical). For both analysis, the independent variables, bank-size measures, are continuous. The following two methods were chosen:

1 **Pearson's correlation**: fits best when analysing continuous independent and dependent variables. The Pearson correlation calculates the dependency between continuous variables bank-size measures and target frequency.

2 **Chi-square**; is a statistical test of independence to determine the dependency of two variables. From the definition we could deduce the application of chi-square technique in feature selection. Suppose there is a target variable (i.e., the class label) and some other features (feature variables) that describe each sample of the data. Then calculate chi-square statistics between every feature variable and the target variable and identify the presence of a relationship between the variables and the target. If the target variable is independent of the feature variable; that future variable could be discarded. If they are dependent, the feature variable is relevant.

Filter methods rank features on specific criteria, but do not consider the relation between variables. Wrappers often provide better results than filters since they are tuned to the specific interaction between an induction algorithm and its training data. The filter method was therefore used as an selection criteria, the wrapper techniques in next section would continue to find the best variables.

F.1.2. Wrappers

The Wrapper method evaluates a subset of variables detecting possible interactions between them. The following wrappers were used in this feature analysis.

1 The **Recursive Feature elimination (RFE)** selection method is chosen to rank features on a recursive process. It uses a greedy optimisation algorithm to find the best performing feature subset. At each iteration the feature importance are measured and the less relevant one will be removed. The recursion is needed since for some measures the relative importance of each feature can change substantially when evaluated over a different subset of features during the step-wise elimination process (in particular for highly correlated features). The inverse order in which features are eliminated was used to construct a final ranking.

RFE selects features by recursively considering smaller sets of features by assigning weights to the features. The weights are the coefficients of the linear model when analysing the continuous dependent variable - target frequency and the coefficients of the logistic regression for the categorical dependent variable - being target. The least important features are pruned from the current set of features. The process is recursively repeated until the desired number of features is selected.

The results showed that RFE was able to identify strong causal variables with a few highly correlated variables, but it did not detect other causal variables. When the number of observations is insufficient (rule of thumb: 50 per category) there is risk of over-fitting. The rule of thumb is to have 50 rows per indicator. In this case 9 features times 50 is 450 were analysed.

F.1.3. Embedded methods

Embedded methods combine the qualities of filter and wrapper methods. They are implemented by algorithms that have their own built-in feature selection methods. Some of the benefits of this method are: they are highly accurate, they generalize better and they can be interpreted.

Embedded methods learn which features best contribute to the accuracy of the model while the model is being created (see figure F.3). The most common type is the regularisation methods also called penalisation methods that introduce additional constraints into the optimisation of a predictive algorithm (such as a regression algorithm) that bias the model toward lower complexity (fewer coefficients).

1 Logistic Regression L1 & Linear Regression L1 (Lasso Regression): Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds “absolute value of magnitude” of coefficient as penalty term to the loss function.

2 Random Forest: consists of 4 –12 hundred decision trees, each of them built over a random extraction of the observations from the dataset and a random extraction of the features. Not every tree sees all the features or all the observations, and this guarantees that the trees are de-correlated and therefore less prone to over-fitting. Each tree is also a sequence of yes-no questions based on a single or a combination of features. At each node (this is at each question), the three divides the dataset into 2 buckets, each of them hosting observations that are more similar among themselves and different from the ones in the other bucket. Therefore, the importance of each feature is derived from how “pure” each of the buckets is.

F.2. Result of the Feature Selection

Feature Selection in the Logistic Regression

For the Logistic Regression Feature Selection, the following feature measures will be used: Chi-square, REF (Logistic Regression), Lasso Regression (Logistic), and Random Forest Classifier (appendix F). All these methods are useful when analysing continuous independent variables given a dichotomous dependent variable. The results can be found in table F.1. The table shows that revenues, loans of customers, equity and branches will represent bank-size in the Logistic Regression. After adding them to the model, the model was again examined on its multi-collinearity. One-factor authentication had to be removed, because the VIF was to high. Now the model was free from multi-collinearity.

Feature	Chi-2	RFE	Logistic	Random Forest	Total
Revenues	False	True	True	True	3
Loans of customers	True	True	True	False	3
Equity	False	True	True	True	3
Branches	True	True	True	False	3
Total assets	True	True	False	False	2
Customers	False	False	True	False	1
Employees	True	False	False	False	1
Deposits	True	False	False	False	1
Online customers	False	False	False	False	0
Net income	False	False	False	False	0
Market capital	False	False	False	False	0

Table F.1: Feature analysis: Logistic Regression

Feature	Logistics	Pearson	RFE	Random Forest	Total
Total assets	True	True	True	False	3
Customers	True	True	True	False	3
Net income	True	True	True	False	3
Deposits	True	True	True	False	3
Loans of customers	False	True	True	False	2
Revenues	False	True	False	False	1
Online customers	False	True	False	False	1
Market capital	False	True	False	False	1
Equity	False	True	False	False	1
Employees	False	True	False	False	1
Branches	False	True	False	False	1

Table F.2: Feature analysis: Negative Binomial Regression

Negative Binomial Regression

For the Negative Binomial Regression Feature Selection, the following feature measures will be used: Pearsons Correlation, REF (Linear Regression), Lasso Regression, and Random Forest Classifier (see Appendix F). These methods are useful methods to analyse a continuous independent variable and a continuous dependent variable. The results can be found in

Table F.2. The table shows that Total assets, Customers, Net income and Deposits of customers are the most important features. However, when adding those variables to the other predictions, Net income indicates a VIF >5 and will to be removed from the list. Loans of customers will be then added. At this point, all VIF 's where below five with the result that Total assets, Number of customers, Deposits of customers and Loans of customers will represent bank-size in the Negative Binomial Regression. By selecting models where VIF < 5, multicollinearity is no longer an issue. Interaction variables are therefore not necessary.

F.3. Conclusion Feature Selection

Feature Selection solved the multicollinearity in the regression model by selecting the most suitable bank-size measures to represent bank-size in both regression analysis, resulting in more stable coefficients. The feature selection analysis identified that **Revenues, Loans of customers, Equity and Branches** will represent 'bank-size' in the Logistic Regression. In the Negative Binomial model this will be **Total assets, Customers, Deposits of customers and Loans of customers**.

G

Principal Components Contribution

Table G.1: Components Contributions (PC1,PC2) of logistic and NB dataset

	Logistic Dataset		**NB Dataset**	
	PC1	PC2	PC1	PC2
Revenues	0.32070738	-0.04298278	0.37852748	-0.002929548
Equity	0.32143593	-0.01305659	0.31313154	0.024066346
Total Assets	0.39937137	-0.01856307	0.39109229	0.020743189
Market Capital	0.07907216	0.**68180549**	0.08105245	-0.086210014
Net Income	0.37210921	0.00249165	0.35442678	-0.018934703
Number of Customer	-0.034410529	0.29784789	0.28648243	0.049774871
Employees	0.33417480	0.01858290	0.32372433	-0.002182890
Branches	0.27313210	-0.03119483	0.26248736	0.068215481
Loans Of Customers	0.29418723	0.03201392	0.28271350	-0.023333978
Deposits of Customers	0.36313044	-0.02816310	0.36133396	0.012001815
Number of Online Customer	-0.02245882	**0.72694698**	-0.02499993	-0.721051364

H

Results regression analysis

H.1. Performance Logistic Regression: (un)balanced dataset

From Table H.1, Figure H.2 and Figure H.1 it is clear that the adjusted balanced dataset using SMOTE performs way better. Moreover, the results from the logistic regression of the unbalanced data shows that bank-size is not significant with being targeted, and other languages are significant compared to the logistic analysis using balanced data. These languages are: Estonian and Danish.

Table H.1: Comparison performance of Regression model using unbalanced and balanced data

Logistic Model	Unbalanced Data	Balanced Data
AUCvalue	0.7279628	0.8180442
McFadden Rsquare	0.1360577	0.3359655

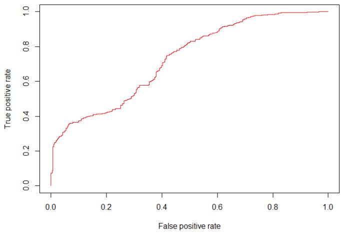


Figure H.1: ROC-curve Unbalanced Logistic Model

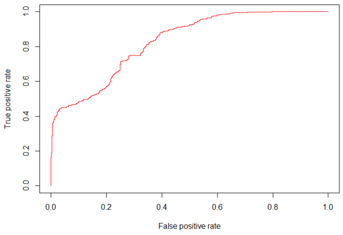


Figure H.2: ROC-curve Logistic Model after SMOTE

H.2. Results Logistic Model

$$
\begin{aligned}
&Call:\\
&\; glm.nb(formula = is\_targeted\ banksize + has\_parent + Rank2017+\\
&\qquad + pop\_score + auth1FA + auth2FA + langEnglish + langGerman+\\
&\qquad langFrench + langDutch + langItalian + langSpanish + langPortugese+\\
&\qquad langGreek + langCzech + langSlovak + langSlovenian + langPolish+\\
&\qquad langHungarian + langRomanian + langBulgarian + langDanish+\\
&\qquad langSwedish + langFinnish + langLatvian + langEstonian+\\
&\qquad langLithuanian + lang\_count, family = "binomial", data = my\_data.new\_logit)
\end{aligned}
\tag{H.1}
$$

Table H.2: Summary Stepwise Logistic_model

	Dependent variable:				
	(1)	(2)	(3)	(4)	(5)
Country_Austria	19.264	19.242	19.613	20.506	22.195
	(123.495)	(123.235)	(197.505)	(191.311)	(175.828)
Country_Belgium	16.241	16.068	16.212	16.649	15.871
	(123.494)	(123.235)	(197.505)	(191.311)	(175.828)
Country_Bulgaria	35.220	35.212	33.706	33.609	35.245
	(270.233)	(270.217)	(1,471.819)	(1,471.000)	(1,469.067)
Country_Croatia	20.270	20.115	21.833	22.286	23.450
	(123.495)	(123.235)	(197.505)	(191.311)	(175.828)
Country_Cyprus	17.954	17.953	18.203	18.101	19.670
	(123.495)	(123.235)	(197.505)	(191.312)	(175.828)
Country_Czechia	22.514	22.355	8.349	7.988	9.884
	(123.499)	(123.239)	(769.937)	(782.612)	(709.311)
Country_Denmark	15.713	15.766	14.957	12.549	9.171
	(123.495)	(123.235)	(6,570.275)	(6,566.241)	(6,577.324)
Country_Estonia	17.229	16.708	18.012	17.903	19.321
	(123.495)	(123.235)	(197.505)	(191.312)	(175.828)
Country_Finland	34.974	34.703	55.558	55.722	60.530
	(586.461)	(534.819)	(6,580.806)	(6,573.821)	(6,585.937)
Country_France	18.649	18.823	19.575	18.805	20.177
	(123.495)	(123.235)	(197.505)	(191.311)	(175.828)
Country_Germany	17.726	17.722	18.739	19.768	21.947
	(123.494)	(123.235)	(197.505)	(191.311)	(175.828)
Country_Greece	20.774	20.816	18.276	18.200	19.105
	(123.499)	(123.239)	(197.510)	(191.317)	(175.834)
Country_Hungary	16.490	16.215	13.237	13.375	11.133
	(123.495)	(123.235)	(197.511)	(191.329)	(175.846)
Country_Ireland	21.793	21.785	23.087	23.498	23.248
	(123.499)	(123.239)	(197.507)	(191.314)	(175.831)
Country_Italy	18.442	18.517	18.385	19.163	18.916
	(123.494)	(123.235)	(197.505)	(191.312)	(175.829)
Country_Latvia	18.130	18.087	19.604	20.213	21.691
	(123.495)	(123.236)	(197.505)	(191.312)	(175.828)
Country_Lithuania	20.855	20.766	4.304	6.206	5.921
	(123.499)	(123.239)	(6,525.628)	(6,525.444)	(6,525.008)
Country_Luxembourg	18.364	18.982	19.464	19.435	21.056
	(123.495)	(123.235)	(197.505)	(191.311)	(175.828)
Country_Malta	0.050	0.191	0.518	0.662	−0.571
	(417.598)	(402.641)	(660.766)	(643.965)	(552.848)
Country_Netherlands	19.768	19.530	20.295	20.358	20.684
	(123.495)	(123.236)	(197.505)	(191.312)	(175.828)
Country_Poland	20.482	20.408	53.134	53.467	53.374
	(123.495)	(123.235)	(824.562)	(834.663)	(759.829)
Country_Portugal	16.292	16.310	13.717	13.445	12.008
	(123.495)	(123.235)	(197.508)	(191.314)	(175.831)
Country_Romania	18.254	18.759	33.923	34.146	34.502
	(123.495)	(123.235)	(367.089)	(360.790)	(330.441)
Country_Slovakia	19.492	19.492	21.315	21.056	22.592
	(123.495)	(123.235)	(1,337.497)	(1,270.908)	(1,085.777)
Country_Slovenia	14.786	14.971	31.461	30.957	28.728
	(123.495)	(123.235)	(367.090)	(360.790)	(330.441)
Country_Spain	16.790	16.771	17.254	17.853	19.593
	(123.494)	(123.235)	(197.507)	(191.314)	(175.834)

Country_Sweden	17.675	17.432	33.972	32.962	31.577
	(123.494)	(123.235)	(789.718)	(755.446)	(846.378)
Country_United.Kingdom	16.220	16.159	17.534	17.376	17.564
	(123.494)	(123.235)	(197.505)	(191.311)	(175.828)
lang_count					
banksize	0.535***	0.504***	0.473***	0.376***	−0.055*
	(0.024)	(0.024)	(0.024)	(0.024)	(0.025)
auth2FA		3.003***	3.273***	3.308***	4.662***
		(0.175)	(0.185)	(0.188)	(0.242)
auth1FA		2.950***	3.164***	3.112***	4.524***
		(0.172)	(0.183)	(0.186)	(0.246)
langEnglish			0.814***	0.800***	0.676***
			(0.089)	(0.094)	(0.119)
langGerman			1.157***	0.767***	−0.347
			(0.181)	(0.182)	(0.225)
langFrench			0.694**	0.880***	0.756**
			(0.227)	(0.235)	(0.280)
langDutch			1.311***	1.310***	1.838***
			(0.197)	(0.201)	(0.266)
langItalian			2.259***	2.126***	1.119
			(0.374)	(0.364)	(0.610)
langSpanish			1.261	1.190	−0.388
			(0.891)	(0.980)	(1.541)
langPortugese			4.306***	4.096***	5.061***
			(1.104)	(1.062)	(1.078)
langGreek			3.954***	4.708***	4.275***
			(1.092)	(1.107)	(1.099)
langCzech			15.276	15.503	15.215
			(744.174)	(758.868)	(687.172)
langSlovak			−0.888	−0.145	−0.284
			(1,322.835)	(1,256.427)	(1,071.446)
langSlovenian			−15.143	−14.725	−11.005
			(309.429)	(305.891)	(279.778)
langPolish			−30.947	−31.121	−30.426
			(800.559)	(812.443)	(739.205)
langHungarian			3.665*	3.796	4.248
			(1.591)	(2.590)	(2.530)
langRomanian			−13.871	−14.561	−13.576
			(309.429)	(305.891)	(279.778)
langBulgarian			34.588	35.703	27.539
			(1,580.313)	(1,576.983)	(1,560.695)
langDanish			2.408	4.380	8.154
			(6,567.305)	(6,563.454)	(6,574.974)
langSwedish			−14.606	−13.014	−10.892
			(764.622)	(730.820)	(827.913)
langFinnish			−3.529	−4.995	−10.469
			(6,567.302)	(6,563.453)	(6,574.972)
langLatvian					
langEstonian					
langLithuanian			17.833	15.928	16.500
			(6,522.639)	(6,522.639)	(6,522.639)
pop_score					5.337***
					(0.139)

has_parent				1.890*** (0.101)	1.205*** (0.144)
Rank2017				5.489*** (0.517)	2.287*** (0.570)
Constant	−17.462 (123.494)	−20.424 (123.235)	−22.833 (197.505)	−23.473 (191.311)	−26.182 (175.828)
Observations	24,376	24,376	24,376	24,376	24,376
Log Likelihood	−13,789.500	−13,512.610	−13,203.350	−12,831.780	−11,219.630
Akaike Inf. Crit.	27,638.990	27,089.230	26,508.710	25,769.550	22,547.260
Note:					*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets

Note: *p<0.05; **p<0.01; ***p<0.001
Standard errors in brackets

H.3. Results Negative Binomial Regression

Call:

$$\begin{aligned}
&glm.nb(formula = id\_attack\_count\ lang\_count + country + threat\_name + \\
&\quad year + unique\_attackurl\_count + banksize + auth1FA + auth2FA + \\
&\quad langEnglish + langGerman + langFrench + langDutch + langItalian + \\
&\quad langSpanish + langPortugese + langGreek + langCzech + langSlovenian + \\
&\quad langPolish + langSlovak + langEstonian + langHungarian + \\
&\quad langRomanian + langBulgarian + langDanish + langSwedish + \\
&\quad langFinnish + langLatvian + langLithuanian + pop\_score + \\
&\quad Rank2017 + has\_parent, data = my\_data.new\_nb, init.theta = 2.743436067, \\
&\quad link = log)
\end{aligned} \tag{H.2}$$

Table H.3: Summary Negative Binomial_model

	Dependent variable:			
	Stepwise			
	(1)	(2)	(3)	(4)
lang_count	0.036*** (0.008)	0.036*** (0.008)	−0.112*** (0.030)	−0.141*** (0.030)
countryBelgium	0.0004 (0.068)	0.072 (0.071)	−0.604*** (0.132)	−0.646*** (0.131)
countryBulgaria	−0.022 (0.066)	−0.026 (0.066)	−0.331 (0.292)	−0.208 (0.293)
countryCroatia	−0.128* (0.065)	−0.040 (0.071)	−0.090 (0.087)	−0.002 (0.087)
countryCyprus	−0.185 (0.158)	−0.204 (0.158)	−0.279 (0.220)	−0.212 (0.217)
countryCzechia	0.827*** (0.071)	0.849*** (0.072)	0.645*** (0.154)	0.548*** (0.154)
countryDenmark	0.791*** (0.090)	0.886*** (0.096)	0.002 (0.653)	−0.554 (0.648)
countryEstonia	−0.323 (0.182)	−0.215 (0.183)	−0.194 (0.192)	−0.345 (0.193)
countryFinland	0.498*** (0.117)	0.532*** (0.117)	0.538*** (0.157)	0.414** (0.157)
countryFrance	0.272*** (0.061)	0.263*** (0.061)	0.078 (0.092)	−0.139 (0.094)
countryGermany	−0.142** (0.046)	−0.161*** (0.046)	−0.166*** (0.050)	−0.024 (0.051)
countryGreece	−0.256* (0.109)	−0.277* (0.109)	−0.449* (0.224)	−0.543* (0.226)

countryHungary	0.591***	0.584***	0.325*	0.112
	(0.100)	(0.100)	(0.136)	(0.136)
countryIreland	0.629***	0.657***	0.380***	0.168
	(0.086)	(0.086)	(0.096)	(0.097)
countryItaly	−0.168***	−0.176***	0.106	−0.074
	(0.050)	(0.050)	(0.118)	(0.118)
countryLatvia	0.241	0.332*	0.395*	0.535**
	(0.161)	(0.164)	(0.175)	(0.174)
countryLithuania	−0.268*	−0.163	0.608	0.786
	(0.135)	(0.138)	(1.191)	(1.187)
countryLuxembourg	0.030	0.080	0.083	0.079
	(0.082)	(0.084)	(0.101)	(0.101)
countryNetherlands	0.454***	0.470***	−0.298*	−0.502***
	(0.073)	(0.073)	(0.131)	(0.131)
countryPoland	0.286***	0.284***	−0.149	−0.396*
	(0.059)	(0.059)	(0.175)	(0.176)
countryPortugal	0.138	0.119	0.927***	0.562***
	(0.096)	(0.096)	(0.138)	(0.141)
countryRomania	0.497***	0.541***	0.515***	0.434***
	(0.077)	(0.078)	(0.098)	(0.099)
countrySlovakia	−0.129	−0.138	−0.264	−0.279
	(0.090)	(0.090)	(0.153)	(0.152)
countrySlovenia	−0.229	−0.250	−0.284	−0.362
	(0.278)	(0.278)	(0.433)	(0.437)
countrySpain	0.256***	0.244***	−0.369**	−0.382***
	(0.071)	(0.071)	(0.115)	(0.114)
countrySweden	0.293***	0.379***	0.157	−0.017
	(0.061)	(0.065)	(0.132)	(0.132)
countryUnited Kingdom	0.091	0.108	−0.118	−0.316***
	(0.059)	(0.060)	(0.073)	(0.074)
threat_nameCitadel	1.313***	1.301***	1.288***	1.327***
	(0.269)	(0.269)	(0.267)	(0.266)
threat_nameCoreBot	−0.197	−0.146	−0.071	0.185
	(0.676)	(0.675)	(0.673)	(0.666)
threat_nameDridex-Loader	2.552***	2.544***	2.532***	2.585***
	(0.268)	(0.267)	(0.265)	(0.264)
threat_nameGootkit	1.632***	1.622***	1.609***	1.654***
	(0.268)	(0.268)	(0.265)	(0.264)
threat_nameGootkitLoader	0.575*	0.567*	0.550*	0.605*
	(0.270)	(0.270)	(0.268)	(0.267)
threat_nameGozi-EQ	1.290***	1.286***	1.279***	1.331***
	(0.274)	(0.274)	(0.272)	(0.271)
threat_nameGozi-ISFB	4.009***	3.999***	3.989***	4.038***
	(0.268)	(0.268)	(0.265)	(0.264)
threat_nameKINS	1.034***	1.018***	0.988***	0.993***
	(0.275)	(0.275)	(0.273)	(0.272)
threat_nameKronos	1.751***	1.739***	1.723***	1.775***
	(0.269)	(0.269)	(0.267)	(0.265)
threat_nameMatrix	−0.568	−0.523	−0.442	−0.202
	(0.717)	(0.717)	(0.713)	(0.711)
threat_nameNuclearBot	0.675*	0.666*	0.639*	0.686**
	(0.269)	(0.269)	(0.267)	(0.266)
threat_nameNymaim	−1.301	−1.198	−1.218	−0.826
	(0.979)	(0.979)	(0.972)	(0.968)
threat_namePkybot	−1.271**	−1.304**	−1.308**	−1.195*
	(0.484)	(0.484)	(0.481)	(0.479)

threat_nameQadars	1.609***	1.600***	1.585***	1.629***
	(0.268)	(0.268)	(0.266)	(0.265)
threat_nameQakbot	−0.603	−0.507	−0.522	−0.435
	(1.208)	(1.208)	(1.202)	(1.199)
threat_nameRamnit	1.192***	1.172***	1.182***	1.201***
	(0.278)	(0.278)	(0.275)	(0.274)
threat_nameRamnit-BankerModule	1.380***	1.351***	1.369***	1.435***
	(0.369)	(0.368)	(0.365)	(0.362)
threat_nameReactorBot	0.494	0.481	0.454	0.407
	(0.611)	(0.612)	(0.605)	(0.601)
threat_nameRetefe-v2	2.789***	2.756***	2.614***	2.644***
	(0.282)	(0.282)	(0.281)	(0.279)
threat_nameTheTrick	3.654***	3.642***	3.543***	3.582***
	(0.268)	(0.268)	(0.266)	(0.265)
threat_nameTinba-v1	1.295***	1.287***	1.269***	1.319***
	(0.272)	(0.272)	(0.269)	(0.268)
threat_nameTinba-v2	0.610*	0.601*	0.584*	0.639*
	(0.270)	(0.270)	(0.268)	(0.267)
threat_nameZeuS	0.649*	0.638*	0.586*	0.621*
	(0.284)	(0.284)	(0.282)	(0.281)
threat_nameZeus-Action	0.109	0.077	0.136	0.176
	(0.363)	(0.363)	(0.360)	(0.359)
threat_nameZeuS-OpenSSL	2.255***	2.244***	2.230***	2.279***
	(0.268)	(0.267)	(0.265)	(0.264)
threat_nameZeus-Panda	2.376***	2.364***	2.355***	2.380***
	(0.268)	(0.268)	(0.266)	(0.265)
year2017	0.060***	0.061***	0.059***	0.064***
	(0.017)	(0.017)	(0.017)	(0.017)
unique_attackurl_count	0.017***	0.017***	0.017***	0.014***
	(0.0004)	(0.0004)	(0.0004)	(0.0005)
banksize	0.041***	0.041***	0.044***	0.029***
	(0.003)	(0.003)	(0.003)	(0.003)
auth1FATrue		−0.033	−0.059	−0.050
		(0.045)	(0.046)	(0.046)
auth2FATrue		−0.146***	−0.153***	−0.144***
		(0.041)	(0.041)	(0.041)
langEnglishTrue			0.167***	0.181***
			(0.040)	(0.040)
langGermanTrue			−0.021	−0.034
			(0.055)	(0.055)
langFrenchTrue			0.178*	0.200**
			(0.077)	(0.077)
langDutchTrue			0.713***	0.713***
			(0.102)	(0.100)
langItalianTrue			−0.314**	−0.266*
			(0.107)	(0.106)
langSpanishTrue			0.875***	0.810***
			(0.098)	(0.098)
langPortugeseTrue			−0.955***	−0.903***
			(0.120)	(0.120)
langGreekTrue			0.120	0.141
			(0.222)	(0.223)
langCzechTrue			0.203	0.168
			(0.142)	(0.141)
langSlovenianTrue			−0.013	−0.018
			(0.335)	(0.337)

langPolishTrue			0.352* (0.172)	0.477** (0.173)
langSlovakTrue			0.042 (0.157)	0.074 (0.156)
langEstonianTrue				
langHungarianTrue			0.370*** (0.110)	0.457*** (0.109)
langRomanianTrue				
langBulgarianTrue			0.229 (0.293)	0.228 (0.293)
langDanishTrue			0.854 (0.645)	1.204 (0.639)
langSwedishTrue			0.176 (0.124)	0.364** (0.124)
langFinnishTrue				
langLatvianTrue				
langLithuanianTrue			−0.787 (1.183)	−0.958 (1.180)
pop_score				0.00000*** (0.00000)
Rank2017				0.001*** (0.0002)
has_parent				0.201*** (0.030)
Constant	−0.741** (0.272)	−0.676* (0.276)	−0.460 (0.277)	−0.582* (0.275)
Observations	12,228	12,228	12,228	12,228
Log Likelihood	−32,374.190	−32,365.130	−32,211.210	−32,114.540
θ	2.585*** (0.042)	2.591*** (0.042)	2.686*** (0.044)	2.743*** (0.045)
Akaike Inf. Crit.	64,862.370	64,848.250	64,574.410	64,387.070
Note:				*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets

H.4. Language Model

Given the result of the logistic and negative binomial model, it was expected that language is related to size and two-factor authentication. To validate our expectations, interaction terms are added. Those interaction terms are not complete, and the model will not used be interpreted the significance of the independent, but it provides insight into the relationship of some language features with domain popularity and size.

In below table, the control variables are removed to provide oversight.

Table H.4: Summary Language model

	Dependent variable:	
	Kit	
	(1)	(2)
banksize	0.029*** (0.003)	0.025*** (0.003)
auth2FATrue	−0.144*** (0.041)	−0.201*** (0.042)
langEnglishTrue	0.181*** (0.040)	0.232*** (0.041)
langGermanTrue	−0.034 (0.055)	−0.094 (0.056)

langFrenchTrue	0.200**	0.276**
	(0.077)	(0.084)
langDutchTrue	0.713***	1.280***
	(0.100)	(0.160)
langItalianTrue	−0.266*	0.346*
	(0.106)	(0.147)
langSpanishTrue	0.810***	0.849***
	(0.098)	(0.101)
langPortugeseTrue	−0.903***	−0.063
	(0.120)	(0.171)
langGreekTrue	0.141	0.145
	(0.223)	(0.222)
langCzechTrue	0.168	−0.089
	(0.141)	(0.190)
langSlovenianTrue	−0.018	−3.325***
	(0.337)	(0.965)
langPolishTrue	0.477**	0.708***
	(0.173)	(0.182)
langSlovakTrue	0.074	−0.079
	(0.156)	(0.197)
langEstonianTrue		
langHungarianTrue	0.457***	1.043***
	(0.109)	(0.291)
langRomanianTrue		
langBulgarianTrue	0.228	0.233
	(0.293)	(0.292)
langDanishTrue	1.204	1.524*
	(0.639)	(0.634)
langSwedishTrue	0.364**	−0.074
	(0.124)	(0.166)
langFinnishTrue		
langLatvianTrue		
langLithuanianTrue	−0.958	−0.966
	(1.180)	(1.176)
pop_score	0.00000***	0.00000***
	(0.00000)	(0.00000)
Rank2017	0.001***	0.001***
	(0.0002)	(0.0002)
has_parent	0.201***	0.220***
	(0.030)	(0.030)
auth2FATrue:langSwedishTrue		0.719***
		(0.152)
lang_count:langHungarianTrue		0.373***
		(0.103)
langHungarianTrue:langEnglishTrue		−2.331***
		(0.285)
langFrenchTrue:langDutchTrue		−0.828***
		(0.203)
langItalianTrue:pop_score		−0.00000***
		(0.00000)
langSpanishTrue:langPortugeseTrue		−0.821***
		(0.228)

pop_score:langPolishTrue		-0.00000^{**} (0.00000)
Constant	-0.582^{*} (0.275)	-0.588^{*} (0.274)
Observations	12,228	12,228
Log Likelihood	−32,114.540	−32,007.500
θ	2.743^{***} (0.045)	2.809^{***} (0.047)
Akaike Inf. Crit.	64,387.070	64,186.990
Note:		*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets

H.5. Regression Models : Bank Size selected by feature selection

In Table H.5, the insignificant results are removed to provide oversight. From this stepwise regression it is visible that especially equity is a very strong positive predictor (coefficient of 381), even more than visibility related banks' features. Revenues and total assets have an negative effect, pointing towards controversial results since equity, total assets and revenues relate to each other (See correlation matrix Figure E.11 in Appendix E). This controversy is presumably created by the interaction of the independent variable. The Table shows that the coefficient of Total Assets changes from positive to negative when the visibility factors are added. The negative effect can thus be explained by the interaction between visibility related factors. The other independent variables show similar behaviour compared to logistic regression analysis in Table H.2.

Logistic Stepwise Regression

Table H.5: Summary Stepwise Logistics model: feature selection bank-size measurements

	Dependent variable:			
	Kit			
	(1)	(2)	(3)	(4)
Revenues	-15.613^{***} (1.002)	-15.418^{***} (0.999)	-14.675^{***} (1.069)	-16.305^{***} (1.161)
Equity	541.841^{***} (23.621)	541.101^{***} (23.635)	405.428^{***} (27.895)	380.948^{***} (27.489)
TotalAssets	30.715^{***} (4.218)	29.754^{***} (4.210)	-12.022^{**} (4.011)	-14.684^{***} (3.955)
auth2FA		0.671^{***} (0.101)	1.173^{***} (0.139)	1.162^{***} (0.141)
pop_score			4.618^{***} (0.128)	4.566^{***} (0.133)
has_parent			1.374^{***} (0.136)	1.207^{***} (0.146)
Rank2017			2.783^{**} (0.933)	2.879^{**} (0.914)
langEnglish				0.488^{***} (0.124)
langFrench				0.710^{**} (0.261)
langDutch				0.643^{**} (0.241)
langItalian				1.548^{*} (0.750)
langPortugese				4.462^{***} (1.143)
langGreek				4.252^{***} (1.100)
langHungarian				3.605^{*} (1.512)

Constant	−21.950 (120.576)	−31.190 (10,785.370)	−23.222 (109.737)	−24.553 (177.391)
Observations	24,376	24,376	24,376	24,376
Log Likelihood	−12,621.240	−12,592.490	−11,137.760	−11,029.980
Akaike Inf. Crit.	25,308.480	25,252.970	22,349.520	22,173.950
Note:			*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets	

NB Stepwise Regression

Again, the control and insignificant independent variable are removed form Table H.6 to get a better insight in the significant variables. The table shows that total assets and the number of customers have both a positive influence on the target frequency, meaning that higher total assets and a higher number of customers is in relation with a higher number of targets. Similar to the logistic regression, the bank-size measures are very strong predictors and the behaviour of the other results stay the same.

Table H.6: Summary Stepwise Negative Binomial model: feature selection bank-size measurements

	Dependent variable:				
	Kit				
	(1)	(2)	(3)	(4)	(5)
langEnglishTrue					0.167*** (0.039)
langGermanTrue					−0.017 (0.053)
langFrenchTrue					0.197** (0.076)
langDutchTrue					0.699*** (0.100)
langItalianTrue					−0.252* (0.105)
langSpanishTrue					0.818*** (0.097)
langPortugeseTrue					−0.896*** (0.120)
langPolishTrue					0.486** (0.172)
langHungarianTrue					0.462*** (0.103)
langSwedishTrue					0.371** (0.123)
langFinnishTrue					0.413** (0.156)
langLatvianTrue					0.536** (0.173)
auth2FATrue			−0.165*** (0.041)	−0.115*** (0.030)	−0.157*** (0.041)
pop_score				0.00000*** (0.00000)	0.00000*** (0.00000)
Rank2017				0.0005** (0.0002)	0.001*** (0.0002)
has_parent				0.216*** (0.030)	0.198*** (0.031)
NumberofCustomer		0.755*** (0.119)	0.771*** (0.120)	0.450*** (0.122)	0.530*** (0.122)
TotalAssets		1.139*** (0.187)	1.153*** (0.187)	1.000*** (0.185)	0.924*** (0.187)
Constant	−0.741**	−125.421***	−128.941***	−136.660***	−131.828***

	(0.272)	(35.133)	(35.103)	(34.861)	(34.490)
Observations	12,228	12,228	12,228	12,228	12,228
Log Likelihood	−32,374.190	−32,339.080	−32,328.390	−32,232.660	−32,095.010
θ	2.585*** (0.042)	2.607*** (0.042)	2.614*** (0.042)	2.667*** (0.043)	2.756*** (0.045)
Akaike Inf. Crit.	64,862.370	64,796.160	64,778.790	64,591.320	64,350.020
Note:					*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets

H.6. Literature

Table H.7: Summary of logistic model and negative binomial model towards different metrics

	Dependent variable:			
	Kit	Rented	Private	All
	(1)	(2)	(3)	(4)
EmployeesOver500	1.024***	−0.476***	0.828***	1.075***
	(0.044)	(0.057)	(0.034)	(0.033)
EmployeesUnder250	−1.082***	−8.641***	−0.929***	−1.092***
	(0.045)	(1.000)	(0.033)	(0.034)
RevenuesUnder31	1.699***	−15.474	−0.188*	1.433***
	(0.075)	(76.239)	(0.075)	(0.057)
Constant	−0.100***	0.530***	0.065***	−0.077***
	(0.023)	(0.015)	(0.016)	(0.017)
Observations	13,788	24,376	24,376	24,376
Log Likelihood	−8,941.559	−12,975.680	−16,322.160	−15,909.200
Akaike Inf. Crit.	17,891.120	25,959.370	32,652.310	31,826.400
Note:				*p<0.05; **p<0.01; ***p<0.001 Standard errors in brackets

I

Protocol and transcriptions of expert interviews

I.1. Interview Protocol

setting-up

1 Do you mind if a record this interview?

2 May I use your name, your institution/company name and your role in the institution/company to mention this as a resource and part of scientific research?

Phase 1: Getting a general perspective of potential factors affecting target selection

1 In your opinion: which factor(s) could explain why certain banks are not/less/more being targeted by banking malware? Could you elaborate on how these factors affect the selection of a specific bank?

2 Did you recognise in the past an decrease/increase in malware activity in your organisation, what is according to your opinion the cause for the decrease/increase? This thesis evaluates if the size of a bank in an important factor in explaining how often a bank is targeted.

3 In your opinion: do you think that the size of a bank is a factor that determines if a bank may be targeted less/more often. Could you elaborate on that?

4 Which of the below mentioned bank-size measures do you think will support an adversary's decision to target a bank?

- The revenues of a bank;
- The total assets of a bank;
- The equity of a bank;
- The net-income of a bank;
- The risk-weighted asset of a bank;
- The number of (online) customers;
- The number of branches in the bank organisation;
- The number of employees;
- The loans of customers of the bank and/or
- The deposits of customers

5 Next to financial factors, also other non-financial characteristics of online banking have been analysed. These factors are: authentication factor applied, and the domain popularity.

A In general, how do you think the language-use on the website could affect the target selection?

B Do you think largely used languages, such as English and German, have any effect on the target frequency?

C Would applying 2-factor authentication would make any difference compared to applying 1-factor authentication?

D Would more popular domain lead to the online banking being more/less/not targeted?

6 Do you think that cybercriminals focus more on financial or non-financial factors? *To clarify: the revenue of a bank is a financial factor whereas using a two-factor authentication as a security measure for online banking is a non-financial factor.*

7 What are your expectations towards the use of banking-trojans by cybercriminals in the future?

Phase 2: Discussing results

1 Do the results reflect your expectations, i.e. which factors surprise you and are expected to be less/more important, could you please elaborate on that?

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

3 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?

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

I.2. Interview with the Security Advisor and Risk Lead at ABN AMRO

Phase 1: Getting a general perspective of potential factors affecting target selection

1 **In your opinion: which factor(s) could explain why certain banks are not/less/more being targeted by banking malware? Could you elaborate on how these factors affect the selection of a specific bank?**

Banks that do not have two-factor authentication in place are more likely to be targeted compared to banks that have the two-factor in place. Besides that, the size and the notoriety of the bank play an important part in listing certain banks in the configuration file.

All Dutch banks have two-factor authentication since they are constantly adapting to new technologies. Dutch banks have become (cyber security) mature, because of the huge malware attacks in 2011. Other countries, such as France and Germany do not all have two-factor authentication in place.

However currently we see malware bypassing two-factor authentication easily and that first five steps of the kill-chain has been automated.

2 **Did you recognise in the past an decrease/increase in malware activity in your organisation, what is according to your opinion the cause for the decrease/increase?**

Web banking malware increased when there were no proper detection mechanisms. A lot of web-injects were visible when adversaries found out that there was no proper detection mechanism. At this point several advanced methods are implemented to detect, malicious malware (mostly web injects), including fraudulent transactions and overlay detection..

3 **In your opinion: do you think that the size of a bank is a factor that determines if a bank may be targeted less/more often. Could you elaborate on that? Which of the below mentioned bank-size measures do you think will support an adversary's decision to target a bank?**

- The revenues of a bank;
- The total assets of a bank;
- The equity of a bank;
- The net-income of a bank;
- The risk-weighted asset of a bank;
- The number of (online) customers;
- The number of branches in the bank organisation;
- The number of employees;
- The loans of customers of the bank and/or
- The deposits of customers

When adversaries use malware they probably choose to target big banks. However, if they use phishing, adversaries look for a large target platform. When adversaries use spear-phishing they use malware to check the customer balance. Criminals would then not know the revenues, equity or net income. They then only identify 'big banks' to target and thus are basically looking at the number of customers. Criminals often transfer money from the savings account to customer deposits because the saving account has in most cases more money compared to the deposits. Still, criminals first try to empty the deposits before they try to access the savings account. Employees would only have an impact with Advanced Persistent Threats (for example when using phishing e-mails).

4 **Next to financial factors, also other non-financial characteristics of online banking have been analysed. These factors are: authentication factor applied, language-use on website, and the domain popularity.**

A **In general, how do you think the language-use on the website could affect the target selection?**

Language-use can play a role when developing an overlay. It would be easier to develop an English overlay. However, it is also developed for Dutch and that is a very difficult language. The network of criminals is very large and there are many criminals who speak and/or write more European languages. Maybe language was a factor in the past, but not today anymore.

B Would more popular domain lead to the online banking being more/less/not targeted?

Yes, popularity will result in a higher likelihood to be targeted. However, Tikkie is popular but is not being targeted by malware, but that is because phishing is easier to deploy.

5 **Do you think that cybercriminals focus more on financial or non-financial factors?** *To clarify: the revenue of a bank is a financial factor whereas using a two-factor authentication as a security measure for online banking is a non-financial factor.*

It shifts based on the success of your security. Once you have the right security measures in place and keep them updates, it will not be beneficial to target those banks.

Money mules are the most important assets for criminals to retrieve their money. With the decrease of the number of offline banks and an increase of digital bank customers, it is easier to create mules. The identity (ID and photo) can be stolen or acquired from the internet and used to open a digital bank account. This lowers the barriers to create mules. Money mules can have a very short lifecycle, i.e. two weeks, making it very difficult for banks to act.

Conclusion: banks in countries where onboarding of digital customers is easy will more likely to be targeted due to the ease of creating money mules.

6 **What are your expectations towards the use of banking-trojans by cybercriminals in the future?**

The last two years it was the expectation that the use of mobile malware would increase. However the old business model is still working (i.e. phishing) and this shift is not visible yet. With the implementation of the PSD2 regulation, whereas two-factor authentication is obligated, this could lead to the shift from web malware to mobile malware. Another option is a combination of these malware attacks .

Phase 2: Discussing results

1 **Do the results reflect your expectations, i.e. which factors surprise you and are expected to be less/more important, could you please elaborate on that?**

At first, the two-factor authentication was surprising. However, when thinking about it. Private malware would not have any issue with two-factor authentication since they are able to bypass it.

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

Many factors are interacting and are highly correlated which perhaps doesn't make the results applicable for one factor only. Identifying which variables correlate (positive or negative) will create more insight. A concern related to the model is that mobile malware is not part of the analyses. Here, language will probably be more relevant since it is easier to copy a webpage compared to building an application.

3 **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?**

The model gives insight and is important for threat intelligence purposes. For a bank it identifies how large it is, how many times they are listed in the configuration file, which languages are the most popular to target etc. It can analyse which factors apply for that specific bank and how vulnerable it is for some factors. To a degree it gives insight in which security measures are required and implemented.

4 **What do you think could make the model better in the future?** Adding mobile malware to the model should be considered for follow-on studies.

I.3. Interview with a Security Analyst and a Senior Advisor Criminal Risk Management at a Dutch national bank.

The interview is held with two employees of a Dutch bank, one is working as a security analyst (SA), and the other one is working as a senior advisor criminal risk (SACR). For privacy purposes, they are not named in the interview but referred to based on their function at the bank.

Phase 1: Getting a general perspective of potential factors affecting target selection

1 **In your opinion: which factor(s) could explain why certain banks are not/less/more being targeted by banking malware? Could you elaborate on how these factors affect the selection of a specific bank?**

SACR: In my opinion it depends on who you want to attack. If the bank is the target it will look at revenue, brand exposure, the number of employees. When the target is the client it might be if the authentication is built, the number of clients and how much money is available.

2 **M.Hoppenreijs: Do you think that the adversary does an extensively search towards these factors during his reconnaissance phase in order to make a consideration? Could you elaborate on this?**

SA: This depends on the goal of the adversary. When one is working with phishing mails, the number of clients should be an important consideration. However, when going after the big money, equity, thus the availability of the money is a more important consideration. Criminals would probably conduct more research when going after the big money compared to working with phishing e-mails. In that case an extensive research is necessary to identify which banks should be attacked.

SACR: According to the criminology three elements are required: 1) an attractive target, 2) absence of supervision and 3) a motivated adversary. First, the motivated adversary is present. Second, the attractiveness of the target differs from what you are after. When you choose for big money, you need a bank with a high availability of capital. Malware for a massive deployment is then a more effective instrument. Third, the absence of supervision. There is supervision for fraud in online banking, but to a certain extent there is also a lack of supervision. An important factor is the brand exposure and awareness of a bank. An adversary from China would not know a Dutch national bank, but a local criminal group would know.

SA: Here plays language a role. A Dutch bank focussing on Dutch Citizens has a Dutch website and all the communication is in Dutch. For the adversary, communicating in a foreign language is more difficult than communication in English.

3 **Did you recognise in the past an decrease/increase in malware activity in your organisation, what is according to your opinion the cause for the decrease/increase?**

SACR: A decrease was visible when we implemented the ‘rule-based’ detection. When the success ratio decreases the number of attacks also decreases. But also, the development of new technologies lead to a decrease in targets.

The Netherlands was one of the first countries with a strong internet connection, so we were the first to be attacked. At this moment the adversaries switch to another continent, such as African banks – it is a waterbed movement. For the bank, it is important not to run fast but to run faster than your neighbour.

SA: Besides that, there is a difference in national and international banks. International banks have to extinguish the fires everywhere, whereas national orientated banks only need to focus on that particular land. For the attackers; they moved to countries which are more likely to be successful.

SACR: Another dependent factor is the structure of the cash-out. If you have a proper rule-based detection/management, transactions to foreign countries are alarming. Thus, if adversaries want to succeed, they need people as their ‘money mules’ to cash-out in that particular country. International banks make many transactions abroad, it is more difficult to recognise frauduleus transactions. An international corporate bank is more important than a small national bank that mainly transfers money locally.

4 **In your opinion: do you think that the size of a bank is a factor that determines if a bank may be targeted less/more often. Could you elaborate on that? Which of the below mentioned bank-size measures do you think will support an adversary’s decision to target a bank?**

- The revenues of a bank;
- The total assets of a bank;
- The equity of a bank;
- The net-income of a bank;
- The risk-weighted asset of a bank;
- The number of (online) customers;
- The number of branches in the bank organisation;
- The number of employees;
- The loans of customers of the bank and/or
- The deposits of customers

SA: Depending on the goal of the adversary, bank size could be an important factor. Phishing means are feasible when one wants to get fast money with low technical skills.

SACR: I think brand awareness is next to bank-size an important factor. A well-known international bank has a higher chance to get targeted by multiple malware. It is easy to copy and use similar configuration files.

SA: It also depends on how much time and money the adversaries have at their disposal. Is the adversary able to write a custom-made malware or will he/she use existing malware that will probably be detected.

SACR: Also, the approach of the attack is important. When using the well-known off-the-shelf products, it could be that some of the configuration files are just being copied.

SACR: If I were a criminal, I would start targeting smaller banks because the chance they have effective security measures in place is low. Also, because banks in Europe and the Netherlands have an obligation to have strong authentication. Personally, I would focus on banks that do not have authentication in place.

5 **Next to financial factors, also other non-financial characteristics of online banking have been analysed. These factors are: authentication factor applied, language-use on website, and the domain popularity.**

SA: International bank with a corporate branch are more vulnerable being targeted.

6 **Do you think that cybercriminals focus more on financial or non-financial factors?** *To clarify: the revenue of a bank is a financial factor whereas using a two-factor authentication as a security measure for online banking is a non-financial factor.*

SACR: Difficult to say, but another potential factor can be the chance of being caught. The selection of victims is mostly based on cost-income ratio. When it is easy to obtain money; income is guaranteed. The threats in the Netherlands shared in meetings is that it is difficult to be successful in getting a lot of money. It depends again on the motivation of the adversaries, going after one big fish or after a lot of smaller fishes?

SA: The awareness of internet-users regarding the risk he/she faces is also important. During holidays there were plenty consumer awareness campaigns. A bank has the responsibility towards clients to inform them about the risks and also to take care of them when they are victimised.

7 **What are your expectations towards the use of banking-trojans by cybercriminals in the future?**

SACR: When looking at historical events, everything appears to be an upward and downwards trends. This also applies for banking malware and at this moment I believe we are at a lower activity level. Phishing and ransomware are still profitable. Malware will probably not increase very soon. However, if the profitability of the attacks decreases and there are no other attack-types, malware could increase again. Another possibility is that they move to other countries or deploy more advanced malware.

SA: Everybody is sharing their information and it is all connected. Malware could in the future more focus on customers data. Adversaries that have data/information about a person and his/her private life could easily blackmail that person. Thus, it can be expected that adversaries move towards a new business model utilising personal data. People will suffer in the future from the consequences of the data leaks. On 'marktplaats.nl' false credits and contracts are sold. This seems to be more profitable compared to stealing money via deposits of customers.

Phase 2: Discussing results

1 **Do the results reflect your expectations, i.e. which factors surprise you and are expected to be less/more important, could you please elaborate on that?**

SACR: I think target selection is very connected with the brand awareness, so I would add that variable.

2 **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?**

SA: It is good to have such a model to create insight. There are many developments in the Netherlands and the model is able to show the focus points related to which security measures should be in place.

3 **What do you think could make the model better in the future? SA**: The motivation is in my opinion the most important factor and it shows how many recourses an actor has. This differs for state actors and normal actors.

SACR:The easiness to laundry the money is an important factor but also if virtual currency is available or the potential setting-up of a money-mules network. This comes with social factors i.e. poverty, drugs abuse and social groups. It will be easier to recruit mules in poor countries. On the other hand, in the Netherlands we also see recruitment via Instagram.

SACR: It is not only the banking malware itself but it is about a whole landscape with all the technical, political and social factors involved. It depends on which measures you take and how to open a bank-account, cash-outs, transaction monitoring and many more social components. If you take action on all those points you can decrease the malware activity.

I.4. Interview with an employee of Threat Fabric

ThreatFabric is a cyber security company helping the financial sector to pro-actively detect known and unknown threats.

Phase 1: Getting a general perspective of potential factors affecting target selection

1 **In your opinion: which factor(s) could explain why certain banks are not/less/more being targeted by banking malware? Could you elaborate on how these factors affect the selection of a specific bank?**

There are a multiple reasons for adversaries to list particular banks in the configuration files.

- Adversaries' copy-and-paste configuration files. This was visible with TheTrick, containing similar configuration files of Zeus and Dyre.
- Criminals will make simple search queries to find relevant targets, i.e. "Biggest Banks in The Netherlands".
- There is more risk for a bank to end up being targeted when having international exposure (less risk if a bank is operating at a local versus international level).
- The crimianal "shitlist" contains names of banks that are difficult to defraud due to their malware detection mechanisms. Those banks on the shitlist are removed from the configuration files.
- Criminals started to target private banks but they then found out that corporate banks had more capital and subsequently tried those. Banks providing different products or services (private, corporate and investment) are more likely to be targeted.
- Banking malware spreads like a virus, it often starts by targeting the big banks of a certain country, and when attack on large banks result in successful fraud, smaller banks in the country start to be targeted too. Security maturity over the whole country is, therefore, an important factor for target selection.
- The criminal underground market also plays a part in target selection. Threat actors will request specific demand on those markets, for example for overlays of banking apps from a certain country.

2 **Did you recognise in the past an decrease/increase in malware activity in your organisation, what is according to your opinion the cause for the decrease/increase?**

It depends on the malware threats. For example, in November 2015 the Dyre Group was arrested leading to no activity from this group. And in December there is always a decrease in malware activity. There is no explanation for this phenomenon. It is expected that the numerous internet buys in December play a role in this.

3 **In your opinion: do you think that the size of a bank is a factor that determines if a bank may be targeted less/more often. Could you elaborate on that? Which of the below mentioned bank-size measures do you think will support an adversary's decision to target a bank?**

- The revenues of a bank;
- The total assets of a bank;
- The equity of a bank;
- The net-income of a bank;
- The risk-weighted asset of a bank;
- The number of (online) customers;
- The number of branches in the bank organisation;

- The number of employees;
- The loans of customers of the bank and/or
- The deposits of customers

Bank-size is one of the factors in target selection, similar to the visibility and the number of offers/products of a bank. Revenues would probably not have impact, actors only hope for low-security on the services provided by banks.

4 **Next to financial factors, also other non-financial characteristics of online banking have been analysed. These factors are: authentication factor applied, language-use on website, and the domain popularity.**

A **Do you think largely used languages, such as English and German, have any effect on the target frequency?**

Language plays a role in target selection, especially for those banks that are national focused and therefore providing only their national language on the website.

B **Would applying 2-factor authentication would make any difference compared to applying 1-factor authentication?**

Yes, but it depends on the two-factor authentication mechanism used, weak two-factor authentication mechanisms even increased the chance of successful attacks. For example, one of the two-factor authentication abused is SMS-based. Some infection campaigns also send fake SMS messages with the bank name as sender, containing malicious links to make victims believe that the bank requires setup of additional security measures.

C **Would more popular domain lead to the online banking being more/less/not targeted?** It would have an impact but lower compared to two-factor authentication and language-use.

I.5. In-depth interview Fox-IT analyst

1 **In your opinion: which factor(s) could explain why certain banks are not/less/more being targeted by banking malware? Could you elaborate on how these factors affect the selection of a specific bank?**
At first, language was a very important factor. For example, there was one bank in a certain country that was the only one being targeted in a certain country. The explanation of only targeting this particular bank was that they provide next to their native language also a English interface. It was for the adversaries less difficult to copy this English interface. Besides, when using soft skills, such as social engineering, it is easier to speak English.
Also, the German language is often being targeted, since the east-European countries have a certain knowledge level of the German language. At this moment, interfaces in all languages have been seen and is not longer a barrier. Adversaries using language services to translate every language they would like to use.

2 **Where are the most malware campaigns come from?**
The most malware campaigns are coming from East-Europe and Russia.

3 **Is there a trend in targeted smaller banks?**
Yes, adversaries try to target smaller banks that does not have their security measures in place, i.e. fraud prevention, rule engine, or intelligence providers. On the other hand, smaller banks use third parties to for internet banking.

4 **Did you recognise in the past a decrease/increase in malware activity in your organisation, what is according to your opinion the cause for the decrease/increase?**
Yes, 2011 there was an enormous peak of online fraud, because of the lack of detection. Big banks lost for million of Euros to online fraud.

Also, money mules influences target frequency. When banks have been successful attacked, they try to search money mules to launder the money. However, also the other way around, when there are a lot of money mules available criminals would try to retrieve financial gains from these banks.

5 **Do you think there is a difference between private, kit and rented malware in target selection?** Kit Malware, such as Zeus can be bought from the the underground forums for 10.000 dollars. They also sell a basic support. They are often individuals that often switch, and mostly are based on which web inject are available. Those attacks come and go. Rented malware is often written and people pay monthly a fee for this services. Private malware is malware written by a certain group , but also uses pieces of kit malware. The service providers is very professional and use

6 **The shitlist contains names of banks that are difficult to target due to their detection. Those banks on the shitlist are removed from the configuration files. This has been seen with the Dyre malware, can you confirm this?**
This list is a global list which is updated when certain banks avoid a successful attack. Sometime banks are removed from the configurations list, but are a few week later again added when the group developed its malware.