Cognitive triaging of phishing attacks

van der Heijden, A.

Award date:
2019

Cognitive Triaging of Phishing Attacks

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science (MSc) in Computer Science and Engineering with specialization in Information Security Technology.

Eindhoven University of Technology
Department of Computer Science & Mathematics

Amber van der Heijden
a.v.d.heijden@student.tue.nl

Eindhoven, December 2018

Supervisors
dr. Luca Allodi
Eindhoven University of Technology

Martijn Docters van Leeuwen
Rabobank

Committee
dr. Luca Allodi
dr. Nicola Zannone
dr. Alexander Serebrenik

Abstract

Despite advancements in contemporary anti-phishing solutions, well-forged phishing emails are still successfully passing detection filters and entering people's mailboxes. These phishing emails are reacted upon by humans, who may or may not fall for the phish depending on the persuasiveness of the message. To heighten persuasiveness, phishing emails exploit several well-studied persuasion strategies grounded in cognitive psychology that can be used for the influence of others. No earlier work has aimed to quantify the presence of these influence tactics in phishing emails, nor focused on integrating these measures into a practical risk-based solution.

In this thesis we aim to develop an effective triaging mechanism to automatically prioritize phishing attacks based on their expected degree of success, relying on data from the anti-phishing division of Rabobank, a large financial organization in The Netherlands. We draw from the cognitive psychology literature to characterize the persuasiveness of phishing emails in terms of six principles that can be exploited for the influence of others, namely: `Reciprocity`, `Consistency`, `Liking`, `Social Proof`, `Authority` and `Scarcity`. We construct a topic model based on supervised machine learning to obtain quantitative measurements of these *cognitive vulnerability triggers* in a collection of over 80,000 emails extracted from Rabobank's phishing abuse inbox, and we obtain event alerts on phishing URL clicks from Rabobank's user session monitoring system as measures of phishing efficacy. We use these quantifications in econometric regression analysis to estimate prediction models for phishing success based on the cognitive features of such attacks.

The results of prediction simulations by four model variations, to assure prediction robustness, are stable, and indicate that a small portion of the attacks can be expected to be up to 2 times as effective as the bulk of incoming attacks. These findings empirically illustrate that an effective and fully quantitative triaging mechanism for phishing success can be put in place by response teams to prioritize remediation efforts, e.g. domain takedowns, by first acting on those attacks that are more likely to collect high response rates from potential victims.

Acknowledgements

My deepest gratitude goes to my supervisor, dr. Luca Allodi, for his encouragement, enthusiasm, and thoughtful insight that have laid the foundation for the completion of this thesis. I could not have wished for better support while working on this challenging project. I also thank my supervisor at Rabobank, Martijn Docters van Leeuwen, for letting me explore this fascinating topic and for allowing me the freedom and flexibility to do my best work.

Additionally, I would like to give regards to this thesis' committee members for their time and their interest in my work. I especially thank dr. Alexander Serebrenik, for giving me a first glimpse of what it means to do research; and dr. Nicola Zannone, for all the friendly advice given as a mentor during my first year.

Finally, I want to express my thanks to my fellow interns at Rabobank, for all the welcome distractions during lunch and coffee breaks; my parents and siblings, for the unconditional trust and support; and of course Timmo, for patience and everything else.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Phishing attacks present a significant threat to organizations and their customers. Detection systems have been proposed based on varying defense mechanisms such as signatures of email behaviour [1], textual email-header features [2], and technical impersonation limitations of attackers [3]. Most of these solutions focus on the detection of phishing domains and emails by means of technical traces in order to prevent phishing attacks to happen in the first place. However, despite significant progress in the detection of phishing emails and websites, many attacks still pass through; for example, well-forged phishing emails are essentially indistinguishable from 'legitimate' emails, and 'spear-phishing' targeted attacks are still likely to make it through the filters [4]. The remaining 'false negatives' are parsed and reacted upon by humans: these are generally customers (that may or may not fall for the phish), and employees that forward 'suspected' phishing emails to the fraud and phishing department of the targeted organization. The result is that most 'large enough' organizations operate a phishing-response team that is flooded daily with thousands or tens of thousands 'potential' phishing emails to address.

Once a phishing email reaches the 'phishing inbox' of the organization, the human behind it must promptly decide which action is appropriate to take (e.g to ignore or to start a takedown action). Unfortunately, with thousands of emails arriving daily and in the absence of an objective and quantitative way or prioritizing response, phishing response teams can only act on a first-comes-first-served basis. This implicitly assumes that all (real) phishing attacks arriving in the organization's inbox are equally urgent to mitigate. Detection mechanisms are often in place to assess whether a phishing email has been successful (e.g. by evaluating the requests for internal resources received by the organization's servers and originating from remote domains), but these suffer from inherent technical limitations with the result that many 'successful' phishing attacks remain unknown

and/or unmitigated for potentially very long periods of time. Importantly, measuring (as opposed to *predicting*) success of phishing attacks can only be useful in *after-the-fact* containment processes (i.e. after users gave out their credentials), as opposed to preventing or limiting the attack as soon as possible. By supplying phishing-response teams with automated assessments on the expected impact of incoming phishing attacks, they can more effectively prioritize remediation efforts, e.g. domain takedowns, by first acting on those attacks that are most likely to be successful.

In this thesis we focus on the development of such prediction models for phishing success and show how such models can be used in an effective triaging mechanism for the prioritization of incoming phishing attacks. To achieve this, we rely on persuasion theories grounded in psychology literature to quantitatively characterize phishing emails based on the psychological factors that account for their effectiveness in generating compliant responses from their intended victims.

1.1 Problem definition

The cognitive psychology literature teaches us that convincing humans in performing actions (such as clicking on a link or giving out one's password) is not an easy feat. Cialdini pioneered the definitions of six '*principles of influence*', namely `Reciprocity`, `Consistency`, `Social Proof`, `Authority`, `Liking`, and `Scarcity` as 'cognitive triggers' that, once engaged, can greatly impact the likelihood of a human's decision to comply or not with what he or she is being requested to perform [5]. For example, `Reciprocity` utilizes people's tendency to return favors from others, and `Social Proof` is based on the finding that in uncertain situations people are accustomed to reference the behaviour of others to determine their own actions. These strategies (each further detailed in Chapter 2) have been investigated in multiple fields, e.g. health [6] and negotiations [7], but are generally known for their popular application in sales and marketing [8]. These principles explain why brands are keen to give away free samples of their products on the street (so people feel obliged to actually buy something in return), and part of why online stores often allow customers to write reviews about their purchases (as public display of engagement with a product will make such a product appear attractive to others as well).

These concepts are unified in psychology literature under the umbrella of the *Elaboration Likelihood Model* [9], which considers the means by which humans can be influenced in making decisions when receiving certain (possibly irrelevant) information as input. This model distinguishes two main routes to human information processing: the central route, and the peripheral route. When information is processed centrally, the message will be subject to elaborate and rational

analysis of its contents, whereas peripheral processing of information relies on a set of general heuristics for quick and subconscious evaluation of the message. These mental shortcuts allow people to make decisions while under high cognitive load and in times of uncertainty, however, precisely these heuristics can be exploited for the effective influence of others.

Although these theories of persuasion originate from studies on the efficacy of marketing schemes, these same principles have been shown to contribute to the success of phishing attacks. By abusing these vulnerabilities, attackers try to persuade their targets into clicking a phishing link and providing sensitive information. Since these cognitive vulnerabilities are so commonly exploited in actual phishing emails, they provide us with a fitting theoretical framework to characterize the expected effectiveness of such attacks based on the specific cognitive triggers they employ.

Earlier works have confirmed the prevalence [10, 11] and efficacy [12, 13, 14] of Cialdini's cognitive vulnerabilities in phishing attacks, however the effort of integrating this knowledge into a practical solution remains unattempted. We address this gap in literature by showcasing a method to evaluate phishing success in operational settings based on a quantification of the cognitive vulnerabilities embedded in phishing emails. Others have focused on the detection of phishing domains and emails by means of technical traces in order to prevent attacks from happening in the first place [1, 2]. Differently, we focus on the evaluation of the potential of the attacks that, despite the countermeasures in place, make it through and need be reacted upon in a timely manner.

1.2 Research approach

To achieve our aims, we employ techniques from natural language processing and econometrics to build a method and estimation process to evaluate cognitive triggers in phishing emails and to build a *cognitive triaging model* of how successful the organization's response team can expect that email to be. We demonstrate empirically that the resulting estimations can effectively help response teams to more efficiently prioritize their responses by addressing first the (few) attacks that are likely to be highly successful, therefore minimizing costs and increasing response effectiveness. This is all the more relevant as outsourced takedown services often contractually enforce upper limits to the number of takedown requests that can be processed.

We extensively analyze more than eighty thousand phishing emails received by the anti-phishing division of Rabobank, a large financial organization in The Netherlands, quantify the 'cognitive vulnerability triggers' embedded in the attacks, and relate them to the number of accesses to the remote phish domain that the anti-phishing division measured. This allows us to empirically derive

a triaging model that, only based on cognitive features of the incoming email, can predict how many 'clicks' can be expected to be generated, therefore helping response teams by effectively prioritizing their remediation efforts. To showcase this, we employ the derived triaging model to predict how many clicks can be expected from the suspicious emails that have remained 'unmeasured' by the organization (because of intrinsic infrastructural limitations).

1.3 Scope definition

This thesis aims at building a principled analysis that explains *why* one can expect a certain phishing email to be successful, as opposed to build a classifier that 'blindly' maps email bodies to attack succes. This aims to a greater replicability of this study's results that goes beyond the specific organization example used in this thesis and the training data we employed. This allows us to draw conclusions not only on the characteristics of successful emails and how to measure it, but also provides useful insights on training and containment actions that can follow the arrival of a suspicious email.

Outside the scope of this thesis. In this thesis we focus on cognitive aspects of email phishing and leave out other characteristics of phishing email bodies such as email structure, HTML/media elements in emails' body, or features of the landing website out of the scope. Because of the specific application scenario (large financial European organization), we also do not aim at providing 'universal' coefficients or estimations that can be readily re-used in arbitrary settings. This is not possible as different application domains (e.g. energy vs financial or governance), customer bases, and organization processes may play a significant role in the model estimations. However, our method only uses data already available to most large-enough organizations, opening our findings to immediate testing and deployment in other settings to fine-tune our estimations to their specific environment. The proposed methodology and model can be easily replicated across organizations for the derivation of the correct estimations in their own settings. Further, we underline again that we are not building a new phishing detection tool, we are building a metric to predict the success of phishing attacks such that one can prioritize, among phishing resources, which one to takedown first.

1.4 Contributions of this thesis

Our contributions can be summarized as follows:

- we perform a comprehensive review of existing literature on cognitive factors of phishing attacks and related phishing measures;
- we provide the first empirical analysis of cognitive vulnerabilities as exploited in the wild by attackers launching phishing attacks;
- we employ a robust measurement methodology to identify cognitive vulnerability triggers in phishing emails, using supervised Latent Dirichlet Allocation, and a set of bootstrapped econometric simulations to build robust estimations of model coefficients and predictions;
- we show empirically the correlation between exploited cognitive factors and spoofed `From:` addresses with an objective evaluation of phishing success;
- we quantitatively show that triaging phishing emails to prioritize remediation action is possible and effective in an operational setting.

These contributions are relevant in practice as our estimations of phishing success can serve as essential evidence required to motivate takedown requests of associated phishing landing pages. Direct feedback from Rabobank indicates their interest in having an operational implementation of our proposed solution in their environment.

This thesis proceeds as follows: Chapter 2 sets the background for this work in both the cognitive psychology and information security literature; Chapter 3 details the employed data and methodology; Chapter 4 and 5 report the exploratory and cognitive analysis of the data respectively. Finally, Chapter 6 provides a discussion on this work's results and Chapter 7 concludes the thesis.

Chapter 2

Background and Related work

In order to get a proper understanding of the topics relevant to the problem treated in this thesis, this chapter provides an overview of relevant concepts and techniques, as well as a discussion of related literature.

2.1 Definition of phishing

Phishing attacks can be highly varied in complexity and scope, and are often executed in combination with other cyber attacks [15, 16]. Because of this, many different definitions of the term phishing have been adopted in literature. Recently, Lastdrager [17] has unified these diverging opinions into a more comprehensive definition of the term that captures the elements common to all types of phishing, stating that:

> "Phishing is a *scalable* act of *deception* whereby *impersonation* is used to obtain *information* from a *target*." [17, p.25]

In the simplest case, a phishing attack involves an attacker sending out messages to a large number of recipients while impersonating a trusted organization, e.g. a bank, an insurance company, or a governmental institution. The intention of the attacker is often to obtain a financial benefit by deceiving the recipients of the message to reveal sensitive information that can be monetized, such as bank account credentials. These types of phishing attacks are broad in scope and, as a consequence, suffer from a very low response rate. Therefore, scalability is a key factor in providing attackers with a return on investment [18].

Various media allow for mass-communication of phishing messages, including SMS [19], VOIP [20], and social media platforms [21], however, due to its cost-effectiveness, email remains the

preferred medium for phishing message distribution [22]; a single computer can send over thousands of emails per day for negligible costs, and by exploiting the power of a *botnet*, a network of malware infected computers that can be remotely controlled, an attacker can even deliver phishing emails to over millions of targets daily. Another explanation for the popularity of the email medium is that current anti-phishing solutions are still unsuccessful in stopping *zero-day phishing attacks*, attacks that have not yet been discovered, from entering users' mailboxes [23]. Therefore, an attacker can still expect to yield a sufficient number of victims, even if the majority of phishing recipients is not affiliated with the impersonated organization [24].

Although many phishing campaigns are high-volume, phishing attacks can be executed on a lower scale as well, as is generally the case in targeted spear-phishing attacks. In a spear-phishing attack, the attacker uses specific knowledge related to the target(s) to construct tailored phishing messages to increase the target's likelihood of responding. This approach is most commonly employed by attackers aiming to obtain information from a specific individual or organization. Due to the high amount of detail, spear-phishing is much more effective than regular phishing, however since this attack approach requires extensive collection of information about the target, it is much less scalable than a regular phishing attack [17].

2.2 Phishing modus operandi

Attackers have learned to adapt their phishing attacks to use the most effective techniques [25]. Although the employed tactics often differ, they generally follow a similar structure. This structure can be characterized in an attack progression model. Attack progression models, also known as kill chains, were first developed in the military as a means to better understand and defend against potential threats. The core idea relies on the application of knowledge about the different attack stages and their corresponding activities for the identification of potential mitigation strategies. Hutchins et al. [26] were first to develop an attack progression model for application in the cyber attack domain. Later, Mundie [27] adapted this cyber attack progression model to account for the specifics of phishing attacks. The resulting 5-stage model is shown in Figure 2.1.

Figure 2.1: Attack progression model of a phishing attack.

Each of the different stages can be characterized by a number of activities:

1. *Research and Open Source Intelligence*

   In the first stage of a phishing attack, an attacker decides on the general objective of the attack. Generally, the goal is to obtain some kind of information, such as personal credentials or company data. After establishing a primary objective, the attacker starts the collection of email addresses of individuals that can be expected to possess the desired information.

2. *Planning and Preparation*

   In order to launch a phishing attack, several things need to be set-up. Most importantly, the attacker will need to construct a phishing message. Typically this message is presented in the form of an email in which the attacker aims to impersonate a trusted organization.

   On the technical side, this requires an attacker to make sure that the message appears to originate from a legitimate and trusted domain. This can be done by either spoofing the genuine email address, or by sending from a fake domain that only appears "trusted". Email spoofing is possible because SMTP [28], the standard protocol for sending email messages, does not authenticate the sender of an email message out-of-the-box. Several anti-spoofing techniques exist that offer additional authentication mechanisms, such as SPF [29], DKIM [30], and DMARC [31], however, mail servers do not always have these countermeasures implemented. Alternatively, an attacker can register their own fake domain to act as the origin of the message. By registering a domain name similar to the real domain name of the impersonated organization, targets can be tricked into believing the phishing email is legitimate. Additionally, phishing emails often request a reader to click a certain link to visit a webpage, which requires the attacker to create and host a phishing website that allows victims to enter their sensitive information.

3. *Phishing Operation*

   When the phishing attack is prepared, the attacker's next step is to release the phishing emails. In order to gain the required capacity for mass-distribution of phishing emails, attackers often rent or create a botnet that can be remotely instructed to send out the phishing emails. Another possibility is for the attacker to hack into a vulnerable web server [32] and use that server to distribute their phishing emails.

4. *Response and Information Capture*

   Once the phishing messages are released upon their targets, all the attacker has to do is wait

for a response. Phishing web pages typically have control panels that alert the attacker when an individual has provided the attacker's desired information [33]. The attacker can then use the control panel to extract the information provided by the victim.

5. *Attack Culmination and Exploitation*

   Once the attacker has obtained the desired information, the obtained data can be monetized. The attacker can either sell the information to other criminals, or choose to use it personally. For example, in a general banking phishing scenario, the attacker could use the stolen credentials to log in to the online banking environment and transfer money from the victim's bank account to another account. Typically, this account will be one of a *money mule* [34, 35]. A money mule is a person who allows criminals to use their personal bank account for depositing and withdrawing money. In this way, attackers can make it exceptionally hard for authorities to ascertain their identity.

2.3 Psychological factors of phishing emails

The general objective of a phishing email is to enforce a target to comply with a request, such as clicking a link to a phishing webpage and providing sensitive credentials. The effectiveness of these attacks significantly relies on how quickly the message can generate the desired response. The reason for this is twofold: since the phishing message is distributed via email, a recipient can study the phishing message by rereading it multiple times. This increases the chance that a target will discover the illegitimacy of the phishing email [36] as it allows for deliberate processing of the email contents, as opposed to fast and subconscious processing [37]. Furthermore, phishing emails and their respective phishing web pages are active for only 23 hours on average before they are taken down [38], which requires attackers to construct phishing messages that are both believable as well as persuasive in order to make their targets respond as swiftly as possible.

Attackers apply several techniques to increase believability of their phishing messages. First of all, they craft their phishing messages to resemble communications of the impersonated organizations as closely as possible [39]. For this purpose they often include forged quality marks, images, and logos from trusted organizations as well as signals of credibility [40]. Furthermore, they highly personalize the context of the phishing messages to reflect the targeted population [13]. Additionally, believability is enhanced by technical measures, such as spoofing the email address, and matching the phishing web page to the web page of the impersonated organization. One novel

way in which attackers aim to increase the believability of their web pages is by using HTTPS instead of HTTP. The recent gain in popularity of this strategy is easily explained, as web users often interpret the meaning of a HTTPS connection as a signal that the web page is "safe" or "legitimate" [41].

Whereas believability mainly concerns the visuals of the phishing email, persuasiveness is most commonly associated with the text content of the email. By applying social engineering techniques, attackers try to persuade their targets into clicking a phishing link and providing their sensitive information. These persuasive techniques work by exploiting fundamental vulnerabilities in human psychology [22] that can be explained by looking at the different decision-making systems in the human brain. Stanovich and West [37] distinguished between System 1 and System 2 thinking. System 1 thinking is responsible for facilitating fast emotional and subconscious decision-making, whereas System 2 thinking is responsible for the slow logical and conscious decision-making. System 1 thinking supplies humans with numerous heuristics for quick decision-making in times of high cognitive load [42]. Cialdini [5] demonstrated how these mental-shortcuts can be exploited for the influence of others and pioneered the definitions of six *cognitive vulnerability triggers*, namely `Reciprocity`, `Consistency`, `Social Proof`, `Authority`, `Liking`, and `Scarcity`. These influence tactics are best known for their application in sales and marketing [8], although they can be used to explain persuasive efforts in different contexts as well [7, 6]. We provide some general definitions for each of the principles and the vulnerabilities they exploit:

1. `Reciprocity`

   This technique relies on the exploitation of people's tendency to feel obliged to repay the favours of others regardless of whether these favours had been solicited or not [5]. The simple act of providing someone with a small, even non-expensive gift, can influence the recipient to become more likely to say yes to the giver's future requests.

2. `Consistency`

   This principle describes how people generally strive to act in accordance with their prior made commitments [5]. For example, an individual that publicly announces his commitment to caring for the environment will be more likely to exhibit behaviour consistent with that commitment (e.g. sorting household waste).

3. `Social Proof`

   This strategy relies on people's tendency to refer the behaviours of others to determine their own behaviour in uncertain situations [5]. This effect becomes even more powerful the more people behave in a certain way, which can also be used to explain psychological phenomena such as the *bystander effect* [43], in which a group of individuals fails to help a person in need due to diffusion of responsibility within the group.

4. `Authority`

   The technique abuses people's tendency to follow the advice of people we perceive as experts and figures of authority [44, 5]. Product commercials and headlines often utilize this influence technique with the expression of phrases such as 'scientists say' and 'research shows' to increase people's perception of the attractiveness of such products.

5. `Liking`

   This strategy relies on people's tendency to be more inclined to say 'yes' to the requests of people they know and like, and perceive as similar to themselves [5]. Compliance is even increased in the case of (detected) false compliments and other forms of false flattery.

6. `Scarcity`

   This principle describes how people are naturally drawn to things that are exclusive and hard to obtain [5], which explains why popular brands commonly release 'limited editions' of products that are only available during a certain time period. Even when individuals had no prior interest in purchasing a specific product, the opportunity presented by limited-time offers makes the associated products or services more desirable.

Although Cialdini's persuasion theories originate from studies on the efficacy of marketing schemes, these same principles have been shown to explain the successfulness of face-to-face social engineering efforts in the real world [45]. Unsurprisingly, a number of studies have shown the popular application of these cognitive triggers by contemporary attackers in the digital world as well. Akbar [10] performed a quantitative analysis on 207 unique English-language phishing emails to identify to what extent Cialdini's triggers are applied in phishing emails. The results show `Authority`, `Scarcity` and `Liking` to be most popular, occurring in 96.1%, 41.1%, and 21.7% of the analyzed set of phishing emails respectively. Less popular were `Consistency`, `Reciprocity` and `Social Proof`, that were each identified in less than 20% of phishing emails. A similar study was performed by Ferreira et al. [11], who analyzed 52 unique English-language phishing

Table 2.1: Definitions and examples of Cialdini's principles of influence in phishing emails

Persuasion principle	Definition [1]	Phishing text example [2]
`Reciprocity`	People's tendency to feel obliged to repay what another person has provided for them. "I do something for you, now you do something for me."	"While we continue to work hard to keep our network secure, we're asking you to help us keep your account safe. If you did not try to access your account please click here."
`Consistency`	People's tendency to behave in a way consistent with past decisions and behaviors. After making a commitment to a certain view, company or product, people will normally act in accordance with those prior made commitments.	"We believe that you read the terms and conditions before using our service, and we ask you to stop all activities that violate these terms and conditions. Click here to unflag your account for suspension."
`Social Proof`	People's tendency to reference the behavior of others in determining their own behavior. In uncertain situations people commonly follow the actions of the majority.	"We are introducing new security features to our services. All customers must get their accounts verified again."
`Authority`	People's tendency to obey to people of status in authoritative positions. Compliance normally follows from the possibility of punishment for not complying with the requests of the figure of authority.	"Best regards, <name> Executive Vice President of <company name>"
`Liking`	People's preference for saying "yes" to the requests of people they know and like. People are programmed to like others who like them back and who are similar to them.	"We hope you enjoy the ease and convenience you'll get with the ability to manage your accounts from almost anywhere you are."
`Scarcity`	People's tendency to assign more value to items and opportunities when their availability is limited. When something is scarce, people are more willing to take immediate action as not to waste the opportunity.	"If your account information is not updated within 48 hours then your ability to access your account will be restricted."

[1] Definitions based on [5].
[2] Examples drawn from anti-phishing database at `http://www.millersmiles.co.uk`.

samples extracted from their personal mailboxes and from examples found on the internet. Their findings suggest `Liking` to be most popularly used, followed distantly by triggers of `Scarcity` and `Authority`. These works confirm the extensive application of these cognitive vulnerability triggers in real-life phishing campaigns, which indicates that Cialdini's principles provide a fitting theoretical framework for the investigation of psychological factors in phishing emails and other digital correspondence as well. Table 2.1 provides more detailed definitions of all six cognitive triggers and gives examples of how these are commonly exploited in phishing attacks.

Most related to our work, several studies have investigated the efficacy of the different cognitive triggers in the context of phishing attacks. In his seminal work, Workman [12] mapped each of Cialdini's persuasion principles to personality characteristics to investigate the relationships between them in a digital context. The study was performed at a large organization in the financial service industry in the United States, and hypothesized positive correlations between the mapped personality characteristics and the efficacy of the persuasion principles employed in email phishing and *pretext*, a social engineering attack generally launched over the phone. The hypotheses were tested by analysis of subjective data on the employees' self-reported measures of personality characteristics and social engineering behaviors, as well as objective data on observations of employee's behaviors in actual phishing and pretext attacks. Five out of six hypotheses were

supported by the study's results, and suggest that higher measures of the studied personality characteristics correspond to higher measures of susceptibility to the related persuasion principle in individuals. A similar study conducted by Lawson et al. [46] additionally found extroversion to be indicative of heightened susceptibility to several of Cialdini's persuasion principles.

A complementary perspective was provided by Wright et al. [13], who performed a field experiment at a United States university to try to explain the relative efficacy of Cialdini's persuasion principles regardless of personality characteristics. Sixty-four phishing emails were constructed to reflect all mutual combinations of the six persuasion principles and each of these emails was sent to 41 randomly selected university students. Out of all 2,624 targeted students, 178 students clicked the link in the phishing email and supplied their university credentials on a fake web page. Logistic regression on the obtained measures suggests that the presence of `Liking`, `Reciprocity`, `Social Proof`, and `Scarcity` increase the likelihood that individuals will respond to phishing attacks. Remarkably, the findings indicate that the presence of `Authority` decreases the likelihood to respond. This is not only in contrast with Cialdini's theory of persuasion, but with other psychological research indicating people's tendency to be obedient towards authority as well [44, 47]. The authors mention that this unexpected result may be explained by the fact that the authority principle is extensively used in contemporary phishing emails, leading to a higher resistance to this form of persuasion [16]. Interestingly, in a similar experiment conducted by Butavicius et al. [14], `Authority` was found to be most effective. In this experiment 121 students from an Australian university were each shown 12 self-fabricated emails of three different types (genuine, phishing and spear-phishing) that employed combinations of different persuasion principles, and were asked to determine the safety of clicking the link in the email.

Despite these work's contributions to identifying the psychological factors that account for the success of phishing attacks, these previous studies have largely been concerned with addressing the prevalence and efficacy of such cognitive vulnerabilities in phishing emails. Differently, in this work we aim to integrate the knowledge obtained into a practical solution to evaluate phishing success in operational settings based on the cognitive vulnerabilities embedded in these attacks.

2.4 Phishing measures

Measures against phishing have been moderately studied in literature. A number of experimental user-studies has been conducted on the impact of client-side detection-assistance tools [48, 49], how people evaluate phishing web pages [40], and how an individual's demographics can impact

susceptibility [50]. Generally, most work has focused on the detection of phishing domains and emails by means of technical traces in order to prevent phishing attacks to happen in the first place. Multiple phishing detection systems have been proposed that are based on varying detection-mechanisms such as signatures of email behaviour [1], textual email-header features [2], and technical impersonation limitations of attackers [3]. A set of research guidelines for design and evaluation choices in the development of phishing detection systems is presented in [51] with the aim of improving the performance of these systems in real-world applications and facilitating unbiased comparison of performance results between systems. In this work however, we focus on the evaluation of the potential of those attacks that, despite the countermeasures in place, make it through and need be reacted upon in a timely manner

Additionally, several studies have proposed methods for the detection of persuasive elements in text. For this purpose, Bhakta and Harris [52], and Sawa et al. [53] used topic blacklists to scan text for instruction statements targeting sensitive resources, and Ding et al. [54] proposed the word-personality mappings to assess the receptiveness of different personality types to the language employed in persuasive texts. However, these approaches require considerable manual effort to operationalize. Differently, we aim to present a mostly automated mechanism (apart from manual labeling of a small number of training documents) for the detection of such cognitive factors.

In all, this thesis extends the literature outlined in this chapter by showcasing a new automated method to quantify the presence of cognitive triggers in phishing emails, and by integrating these measures into a fully quantitative and risk-based solution for the efficient triaging of incoming phishing attacks that can be operationalized in practice.

Chapter 3

Data collection and Methodology

We previously explained how the efficacy of phishing attacks can be explained by looking at how attackers abuse cognitive vulnerabilities in the human decision-making system. We intend to extend existing literature on this topic by showcasing an approach to integrate measurements of cognitive factors relevant to phishing effectiveness into a practical risk-based solution to help phishing incident response teams in prioritizing their responses more effectively.

3.1 Research objectives and methodology

For the purpose of achieving this goal, we take an empirical approach in addressing the different aims of the study:

> **Objective 1**
>
> To develop a robust method to detect, and *quantitatively measure*, the presence of cognitive vulnerability triggers in phishing attacks.

We consider methods grounded in natural language processing research, and use machine learning to construct a statistical topic model based on a manual analysis of cognitive vulnerabilities in phishing attacks. Such a topic model provides us with a way to uncover from our phishing text corpus for each cognitive vulnerability, the set of words that best represents the presence of the topic in the full corpus. Based on the underlying word-topic distributions, such a topic model can discover the different cognitive vulnerabilities that are represented in a given email. Most importantly, we describe how such a model can be applied to unseen phishing attacks to identify frequencies of cognitive vulnerability triggers present therein.

Our second objective relates to our aim to use these quantitative measurements towards estimating the risk associated with individual phishing attacks. More specifically:

Objective 2

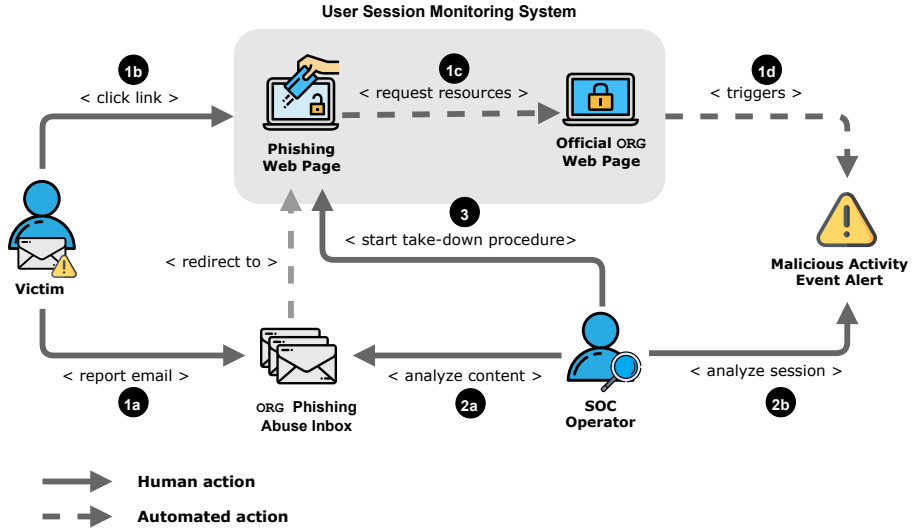
To show how the observed cognitive effects can be integrated into a statistical model to predict the expected efficacy of phishing attacks.

We apply econometric regression analysis to estimate coefficients for multi-variable regression models describing the effect of different vulnerability triggers on phishing attack efficacy. This type of analysis allows us to explore the relationships between the different variables and to infer whether the observed effects can be expected to be present in the larger population as well. Moreover, we quantitatively illustrate how predictions using such a regression model are effective in building a triaging mechanism for phishing incident response in an operational setting.

3.2 Data collection and sanitization

Our analysis relies on a unique dataset from a large phishing email database provided by Rabobank, a large financial organization in The Netherlands with over 8M customers and a multi-billion Euro turnover. Rabobank customers that suspect they have received a phishing email in their personal email accounts are instructed by the organization to forward these emails to an internal Rabobank functional mailbox. In parallel, Rabobank's phishing response team runs a service to detect phishing domains (not necessarily linked with the received phishing emails) by means of internal heuristics and limited to external domains requesting resources internal to Rabobank (e.g. images, forms, logos, CSS files/javascript, etc.). Through this mechanism Rabobank can detect the number of visits to the detected domains by accounting for the unique sessions opened between the (rogue) external and the (legitimate) internal services. Access to this data allows us to perform a rich analysis of the arrival of phishing emails, their characteristics, and to evaluate how often users have accessed malicious domains linked to phishing emails as a proxy measure of 'phishing success'. Figure 3.1 depicts Rabobank's internal process to handle suspect phishing emails.

Overall, we extracted $82,679$ emails that entered this mailbox between February 1st, 2018 and July 31st, 2018 and an additional $5,600$ emails in the period 01 September 2018 and 31 October 2018. Additionally during this latter period and for two weeks after, we collected $7,437$ event alerts for suspicious resource requests from Rabobank's user session monitoring system. A full overview of collected data is included as Table 3.1. The table shows for each organization targeted



SOC Operators collect evidence on the maliciousness of the web domain under investigation such that an external party can perform the notice and take-down requests for the malicious domains.

Figure 3.1: Overview of phishing-related activities at Rabobank

by emails in the dataset how many samples are suspicious and unique, and displays the number of unique phishing origin and victim email addresses identified. Rabobank is indicated in this table as target 'ORG13'. As expected, only a small fraction of obtained emails is targeted at organizations different from Rabobank. Across all targeted organizations we find a large fraction of samples to be 'suspicious' (i.e. the email contains a link to suspicious web domains). Furthermore we note the presence of many duplicates in the obtained sample (duplicate detection methodology discussed in Section 3.2.5) as many clients receive and forward the same phishing emails to the abuse inbox. The time gap in the collection is caused by infrastructural limitations at Rabobank. Due to this we use the whole data collection for descriptive statistics (Chapter 4), and the sample from Sept-Oct 2018 for phishing classification and detection (Chapter 5).

Data limitations. From the data structure, the link between a clicked URL and the specific email from which that click generated is not explicit and can only be reconstructed by exact matching of the destination URL. The reported URLs are generated by an undisclosed set of heuristics and only include remote URLs that request resources internal to Rabobank (e.g. media content stored on Rabobank's servers). This has the effect of limiting the scope of this study to the comparison of the effectiveness of cognitive influence techniques between phishing emails that are likely to have generated the click (as we cannot fully reproduce the process generating the detection of URLs that *could* have been clicked, but have not). This also limits the number of

Table 3.1: Overview of phishing email collections

The column `Samples` reports the total number of entries for that organization. All other columns report number of unique matches for each variable. Susp. indicates a suspicious email. Victims and phishing addresses are measured as the number of unique email addresses in the dataset. Empty cell indicates no measurements in that period for that organization.

	February - July					September - October				
Target	Samples	Susp.	#victims	#phish addr.	Emails	Samples	Susp.	#victims	#phish addr.	Emails
NONE	7952	4740	3250	3335	3545	1505	1391	964	590	550
ORG2	764	661	483	287	126	32	30	26	14	11
ORG3	64	60	38	14	5					
ORG4	1	0	1	1	1	2	2	1	1	1
ORG5	2	2	2	2	2					
ORG6	88	84	67	30	22	2	2	2	2	2
ORG7	494	3	476	20	13					
ORG8	361	293	233	112	69	275	243	140	54	56
ORG9	3074	2686	2014	605	210	235	216	166	76	35
ORG10	62	61	37	8	7	5	5	5	5	4
ORG11	75	58	46	30	21	12	10	10	4	3
ORG12	2	2	2	2	2					
ORG13	69522	60759	38776	1629	1345	3515	2999	2496	195	268
ORG14	28	25	27	5	2	13	13	13	4	2
ORG15	134	118	79	29	20	4	4	4	4	4
ORG16	54	50	46	20	6					
ORG17	2	1	1	1	1					
Sum	82679	69603	45578	6130	5397	5600	4915	3827	949	936

matches between URLs reported in event alerts and URLs linked in emails: redirecting mechanisms in the URLs and domain generation algorithms employed by attackers significantly affect the relative fraction of matches we can find in emails. We compensate for this by means of the analysis methodology that explicitly accounts for the low incidence rate of URL matches.

3.2.1 Email preprocessing

We recursively searched through each raw email message to find header matches of the first original email that arrived in the user's inbox, and extract information on `From`, `To`, `Date`, and `Subject` values. The last separate message body, which should contain the suspect phishing message contents, was then taken for further processing.

Our email dataset contains many mobile text messages resulting from forwarding by a related banking service to the functional mailbox. As they are irrelevant to our research, we discard them by rejecting any email that has a character length below 160 characters, which is the maximum length of a mobile text message. From Figure 3.2 we find the average email body to be of longer length, which indicates that the removal of these short emails should not negatively impact the quality of our data. A manual check of the data suggests that this is indeed the case.

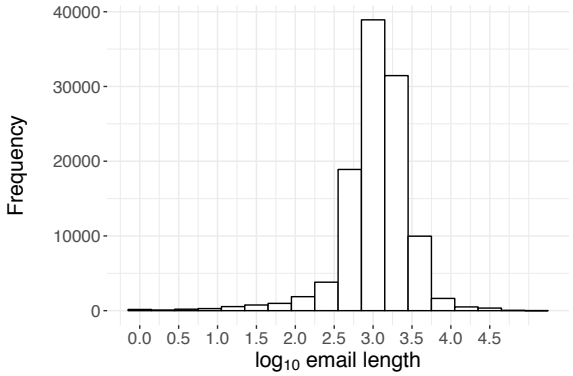


Figure 3.2: Histogram of length of email bodies in original dataset

3.2.2 Target detection

The email address to the mailbox the dataset was extracted from is publicly displayed on the Rabobank website which results in the presence of many unrelated email messages, such as spam advertisements, customer inquiries, and phishing emails targeted at unrelated organizations. To reduce the noise introduced by these 'junk' emails, we identify targeted organizations by a string search operation within email bodies for the names of the most prominent financial organizations in The Netherlands [55]. Subsequently, we extend this list with the names of organizations discovered by manual inspection of a random sample of 200 untargeted emails for better coverage.

3.2.3 Presence of suspicious links

To assess whether an email that is perceived as phishing is likely to constitute an actual phishing email, we check the presence of suspicious links that would not normally appear in a legitimate email originating from Rabobank. As we are only interested in clicks plausibly generated by phishing emails, we look for suspicious links defined as those that direct to any domain that does not belong to Rabobank. We exclude from the heuristic general-purpose domains (e.g. `youtube.com`). Based on this classification we flag emails that contain at least one suspicious link as `Suspicious`, whereas the remaining ones are considered uninteresting within our scope (as we can neither count nor estimate clicks for URLs that do not exist).

3.2.4 Landing webpage extraction

Malicious links embedded in phishing emails often follow several redirects before they arrive at the final phishing landing webpage of which the URL corresponds to the reported URLs in our obtained click event alerts. In order to trace the redirect path from email link to final landing page, we perform HTTP(S) requests for the email embedded URL at collection time and record all intermediary destinations in the history of the final response. Afterwards, we check for full matches (subdomain.second-level-domain.top-level-domain/resource-path) between the URLs from the click event alerts and the URLs obtained from the emails at collection time to create a mapping between clicked URLs and the emails that supposedly generated those clicks.

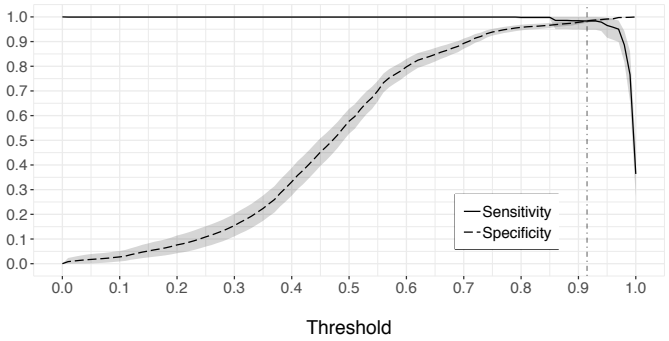
3.2.5 Duplicate detection

One complexity of the unstructured dataset is the possible occurrence of multiple duplicates of the same suspect phishing email. Although the general text content of these duplicate emails is the same, they can still contain slight differences, for instance because of the presence of a recipient's name in the salutation of an email. In order to detect, and subsequently remove, as many of these duplicate emails as possible, we used a fuzzy string matching approach to determine the pairwise similarity for each of the suspect emails in our dataset.

Bag-of-words text representation

Text documents can be represented in different ways. In many cases, the standard string representation of text suffices, however for many classification problems, a vector-space model representation of the text is desirable [56]. One simple and commonly used vector space model in natural language processing is the *bag-of-words* model [57]. For each document, the frequency of each unique word in the document is recorded, such that it can be stored in a word-by-document matrix where each row contains the term frequencies for one unique vocabulary word, and each column is a document vector. More specifically, if we take $E = \{e_1, ..., e_n\}$ to represent our set of emails, and $T = \{t_1, ..., t_m\}$ to represent our set of unique words occurring in E, then each email can be represented as an m-dimensional vector $\vec{v_{e_n}} = (tf(t_1, e_n), ..., tf(t_m, e_n))$ where $tf(t, e)$ is defined by the frequency of term $t \in T$ in email $e \in E$.

We used Python's `scikit-learn` [58] machine learning module to build such a word-by-document matrix containing the term frequency values for all suspect phishing emails in our dataset. As an additional pre-processing step all input was cleaned by removing special characters,



The optimal threshold was found at 0.91 based on the intersection of the mean sensitivity and specificity metrics at all decimal thresholds in $[0, 1]$ across $10,000$ bootstrap simulations with sample size 300.

Figure 3.3: Simulated optimal cosine similarity threshold for duplicate detection

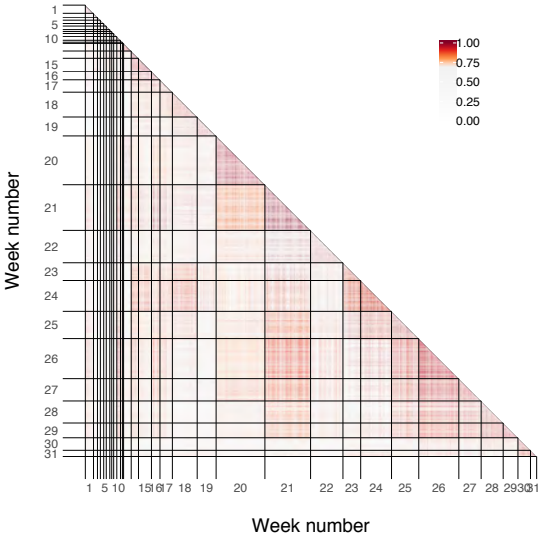
urls, email addresses and line breaks from the text. Furthermore, we applied L^2 normalization [59] to the term frequencies to limit the impact of differences in email lengths such that we can consider the frequency of words relative to each other regardless of total word count, which is found to lead to better performance in solving similarity tasks [60].

The *cosine similarity* can be used as a measure to score the similarity of such normalized vectors by calculating their inner product [61]. This similarity measure expresses the similarity between two vectors in terms of the cosine of the angle between the two vectors and results in a score between $[0, 1]$ where 0 constitutes low textual similarity, and 1 indicates identical messages.

Bootstrap analysis for optimal threshold selection

After computation of the full pairwise similarity matrix for all suspect emails in our dataset, a threshold value was used to determine the lower-bound for the similarity score of emails we consider to be duplicates. In order to determine the most optimal threshold value for our specific dataset we performed a *bootstrap analysis* [62]. A bootstrap analysis involves repeatedly running simulations on samples drawn with replacement from an original sample set in order to estimate statistics on a larger population. The underlying observation is that, at the only condition that the original sample is drawn randomly from the population, a sufficiently large number of samples will provide statistics representative of the full population. A fitting solution for problems concerning dataset of large sizes like ours, which do not generally allow for efficient derivation of the full set of results that qualify as *"ground-truth"*.

We started our bootstrap analysis with a random sample of 300 suspect phishing emails for which manual assessments were made of all pairwise similarities to test the performance of the similarity algorithm across different thresholds. Then, we repeatedly ($n = 10,000$) drew samples



For visualization purposes we report random samples per week of 10% of the emails received in that week. Red represent high similaries above the threshold. We do not observe specific cycles of similar emails, suggesting that any sufficiently long period of time (3-4 weeks) would cover a diverse set of phishing attacks.

Figure 3.4: Pair-wise cosine similarity between email samples

with replacement of size 300 from our manually classified sample and computed the pairwise cosine similarity matrix for all decimal thresholds in the interval [0,1]. For each combination of bootstrap sample and threshold value we computed performance using sensitivity (true positive rate) and specificity metrics (true negative rate) on the resulting confusion matrices according to the following equations:

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{3.1}$$

$$\text{Specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \tag{3.2}$$

A high sensitivity score refers to a high probability of duplicate detection, measured by the proportion of actual duplicates that are correctly identified as being similar, whereas a high specificity score refers to a high probability of non-duplicate rejection, measured by the proportion of actual non-duplicates that are correctly identified as not being similar. The intersections of the mean results for these two performance measures indicate that 0.91 is the optimal threshold value for our dataset, as is visualized more elaborately in Figure 3.3. The narrow confidence intervals show a high stability in the computed measures across all thresholds, indicating a high confidence in the selection of the optimal threshold level.

We use this threshold to calculate the pairwise similarity matrix for a random sample of emails in our dataset. Figure 3.4 reports the similarity scores between emails received during the

observation period for the purpose of visualisation. Dark red indicates high similarity (above the defined threshold). We do not observe specific and systematic cycles of campaigns emerging with repeating patterns across several weeks. This also suggests that any sufficiently long observation period (in the order of 3-4 weeks) may suffice to collect a relatively diverse set of attacks.

3.3 Measurement of vulnerability trigger frequencies

To quantify the presence of cognitive vulnerability triggers in email bodies we construct a supervised topic model based on Labeled LDA [63] (LLDA). LLDA models each input document as a mixture of topics inferred from labeled input data and outputs probabilistic estimates of label-document distributions, i.e $P(label_t|document_m)$, and word counts of label-specific triggers for each input document. In our application the labels correspond to the six cognitive vulnerabilities detailed in Table 2.1, whereas documents correspond to the email bodies. For model training, we randomly sampled 61 suspicious emails and 38 emails linked to at least 1 'click' as reported by Rabobank's telemetry.[1] We manually label them for presence of vulnerability triggers and we use this labeled input data in 5 times repeated 5-fold cross validation to assess performance of our LLDA classification. In k-fold cross-validation [64] the original sample is randomly partitioned into k equal sized parts. Then in each of k runs, one of the subsamples is held out as validation data for the model trained on the other $k-1$ subsamples. In this way, each of the subsamples is used exactly once as the validation data. By repeating this 5-fold process five times, each time resulting in different splits of the data, we obtain more robust model estimations.

Numerous approaches exist to evaluate the performance of multilabel classification problems like ours. Following [65], we consider our problem as a label-pivoted binary classification problem, where the aim is to generate for each label strict yes/no predictions based on the document ranking for that label. For each label, we sort on the per document prediction values, and use the `PROPORTIONAL` method [65, 66] to define a rank-cutoff value that determines the top N ranked items that will receive a positive prediction. For each label, we set $TOPN_i$ equal to the expected number of positive predictions based on training-data frequencies: For label l_i, $TOPN_i = ceil\left(\frac{N^d_{\text{test}}}{N^d_{\text{train}}} * N_i^{\text{train}}\right)$ where N^d_{train} and N^d_{test} refer to the total number of training and testing documents and N_i^{train} is the number of training documents assigned label l_i.

[1] This was done to assure that both 'clicked' emails as well as only 'suspicious' emails were represented in the sample.

Table 3.2: Topic model performance results

We perform LLDA using Gibbs sampling iterations for parameter estimation and inference initialised with hyper parameters $\alpha = 1.0$, $\beta = 0.001$, $k_{\text{labels}} = 6$ and $N_{\text{iterations}} = 1000$.

	Macro (sd)	Micro (sd)
Sensitivity	0.709 (±0.016)	0.807 (±0.016)
Specificity	0.714 (±0.042)	0.813 (±0.038)
Precision	0.718 (±0.025)	0.755 (±0.024)
F1	0.725 (±0.020)	0.760 (±0.020)

We have aggregated the performance results of our topic model using the `PROPORTIONAL` rank-cutoff method in Table 3.2. Unlike other rank-cutoff methods, this approach relies solely on labeling information from the training set, which makes it appropriate for use in real-world production settings as well. We report both macro scores (averages over the individual test scores for each item), and micro scores (computed from the sum of all individual confusion matrices for each item). Macro scores weight each item equally, whereas micro scores give more weight to items with more frequent labels. Our efforts of fine-tuning of the hyper parameters α and β revealed no evident effect on the performance outcomes. Following common heuristics for selection of these hyper parameters, we kept $\alpha = 1.0$ at a relatively high value, indicating that each email is likely to exhibit a mixture of multiple vulnerabilities as opposed to one specific vulnerability, and likewise kept $\beta = 0.001$, to indicate that each of the cognitive vulnerabilities is likely characterized by a smaller collection of specific words, and not by most of the words in the corpus.

The obtained scores indicate a satisfactory fit. A manual analysis on randomly sampled emails from the corpora confirms that the procedure appropriately assigns 'topics' to emails. To derive measures of vulnerability triggers in each email we use the topic-document-word association derived by the LLDA procedure to evaluate the number of words in each document strongly associated to a topic (i.e. cognitive vulnerability). The final model is trained on the complete set of 99 labeled training documents that were used in cross-validation, and then applied to the unseen and unlabeled remainder of the full dataset. In all cases, standard text cleaning procedures have been applied, i.e. special character removal, sentence tokenization, stop-word removal and stemming.

Chapter 4

Exploratory analysis

In this section we discuss the results of an exploratory analysis on the obtained email data set after preprocessing. Table 4.1 reports summary statistics of the collected dataset. For each of the factor variables we report the number of levels of the factor (each discussed later in this chapter), and for each numeric variable we report descriptive statistics.

We find that the maximum values reported for both `Length` (114288 characters) as well as `Liking` (8438 triggers) are disproportionate to the values that we would realistically expect. Manual inspection of these outliers indicates that these values originate from 'junk' emails that were unintentionally not rejected by our data cleaning steps. We observe more realistic values for other cognitive vulnerabilities, e.g. `Scarcity` triggers appear with frequencies between 0 and 189 with the average email containing 39 triggers, and `Consistency` with frequencies between 0 and 176 with the average email containing 27 triggers. Furthermore `Spoof dist.`, indicating the number of character removals/insertions/replacements for a `From` domain to transform into Rabobank's domain name, shows sensible extremes between 0 (indicating a perfectly spoofed domain name) and 50 (extreme dissimilarity between domains), with a mean distance of 8 operations.

Based on Figure 4.2 we evaluate the distribution of reported emails per targeted organization. Unsurprisingly, the vast majority of emails is targeted towards one organization only (Rabobank, $n = 73,673$ or 83%), whereas the remaining 17% of emails target different organizations. Therefore, for the remainder of this report we will only consider emails targeting Rabobank. Additionally, Figure 4.1 shows that the overwhelming majority (92%) of emails are written in the native language of the country in which Rabobank resides, which suggests that attackers invest resources into employing preferred communication language of their victims and target organizations in their emails, which is in line with previous reports on the languages employed in phishing emails [67].

Table 4.1: Descriptive statistics of the collected dataset

The column `type` indicates whether the variable is a factor (f) or numeric (n). The column `lvls` reports number of levels for factors. We do not report summary statistics for factors; details on those are given in Chapter 4. The standard deviation for variable `Date` is reported in days. Estimations of Vulnerability triggers per email is detailed in Chapter 5.

	Variable	type	lvls	Min	0.025q	Mean	Median	0.975q	Max	sd
	Language	f	3							
	Target	f	17							
	To	f	45501							
	From	f	6828							
	Suspicious	f	2							
	Detected	f	2							
	Date	n		2018-02-02	2018-02-20	2018-06-03	2018-05-31	2018-10-05	2018-10-30	50.2
	Length	n		0	268	1609.4	1089	4428	114288	2264.1
Vuln. triggers	Recip rocity	n		0	0	8.8	3	63	330	16.4
Vuln. triggers	Consistency	n		0	0	26.7	13	87	176	27.7
Vuln. triggers	Social Proof	n		0	0	3.3	2	17	90	5.2
Vuln. triggers	Authority	n		0	0	9.7	5	52	121	13
Vuln. triggers	Liking	n		0	0	2.5	0	14	8438	36.3
Vuln. triggers	Scarcity	n		0	0	39.3	37	106	189	28.8
Vuln. triggers	Clicks	n		0	0	58.3	6	336	336	97.9
Vuln. triggers	Spoof dist.	n		0	0	7.7	7	23	50	5.3

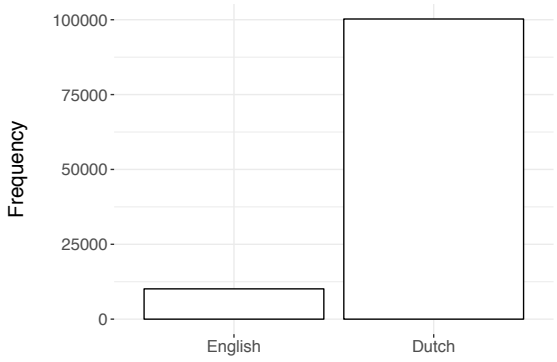


Figure 4.1: Distribution of email languages

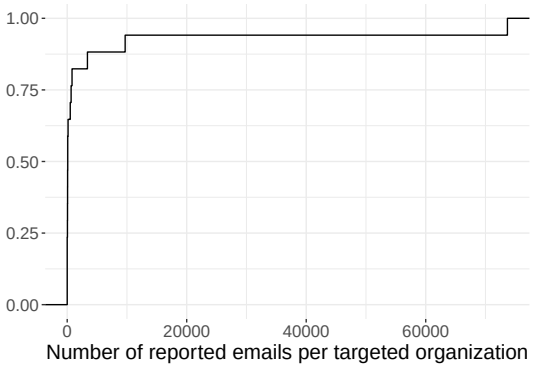
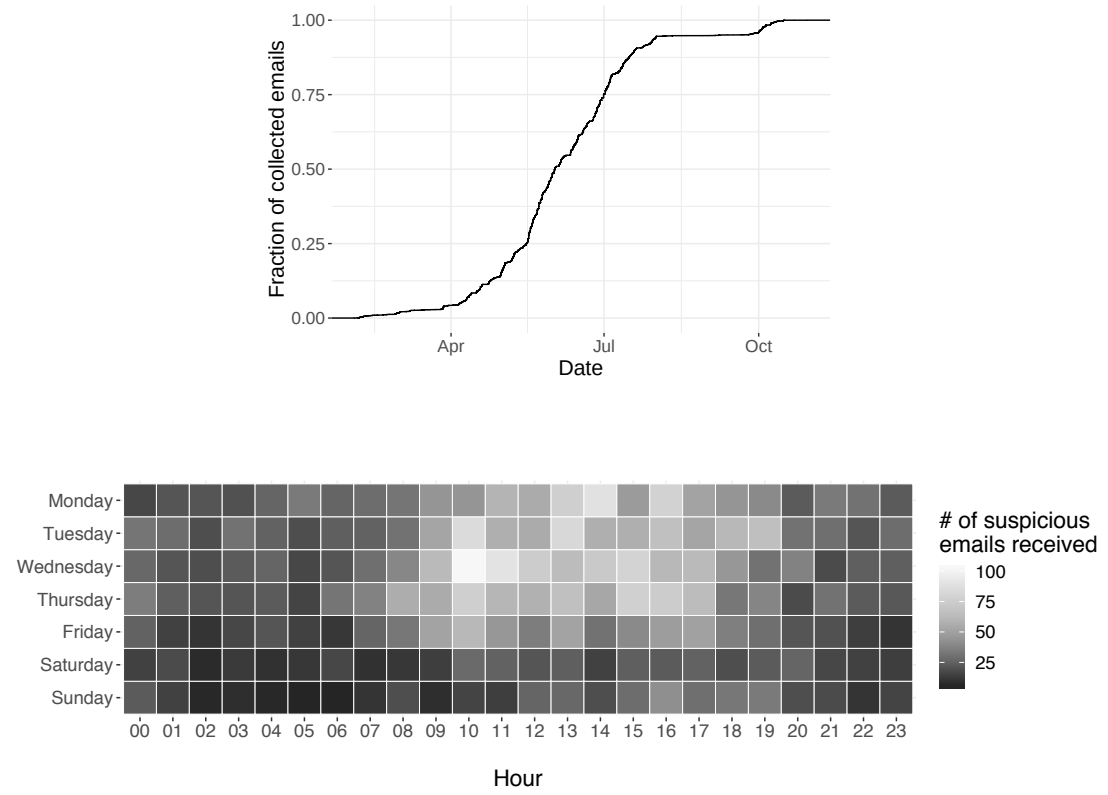


Figure 4.2: CDF of targeted organizations

4.1 Time distribution

Figure 4.3 reports the CDF distribution of email arrivals to Rabobank's phishing inbox. We observe a steady arrival rate through April and the first cutoff date in July 2018, suggesting that email arrival is approximately constant and uniformly distributed in time. As per the arrival of emails in user's inboxes, in Figure 4.4, we can observe that few suspicious emails arrive in the user's inboxes during the weekend, and that activity gradually declines during the work week. We find most emails are distributed during business hours between 9am and 5pm (UTC+1), with the maximum between 9am and 11am. These findings agree with optimal email send days and times for newsletters as reported by analyses from multiple online email marketing services [68, 69, 70, 71].



Daily activity is indicated per hour according to timezone UTC+1. Business hours from Monday through Thursday appear to be the most popular times for distribution of suspicious emails. Significantly less activity can be observed during the weekends and nightly hours.

Figure 4.4: Temporal overview of arrival of suspicious email in users' inboxes by day of the week

4.2 Spoofing and victimization

An email is classified as spoofed based on the Levenshtein distance [72] of the (spoofed) domain from the name of Rabobank. This captures extract string matches as well as small variations that may remain undetected by the user [73] (e.g. `org1.de` → `0rg1.de`). Figure 4.5 depicts the distribution of suspicious and non suspicious emails targeting Rabobank reported by recipients. The CDF is on a log scale to better represent the distribution's log tail. The vast majority of users report only one email, with the almost totality reporting less than 10 emails. This suggests that the distribution of phishing emails is uniform across victims, as is generally the case with untargeted phishing attacks [4, 41]. Only 122 addresses out of 45 thousand report more than 10 emails, and only nine report more than 100 emails.

Interestingly, we find that users are as likely to report emails with suspicious links as emails with no suspicious link, which indicates that it may be hard for users to distinguish between phishing emails and their actually legitimate counterparts. Figure 4.6 reports the distribution of

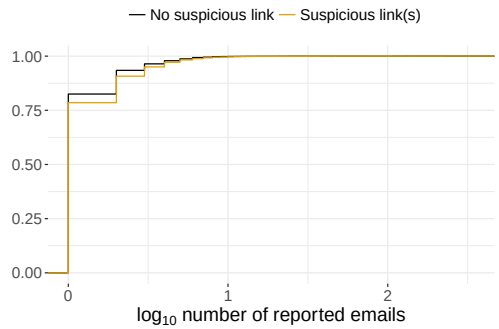


Figure 4.5: CDF of emails reported by victim addresses

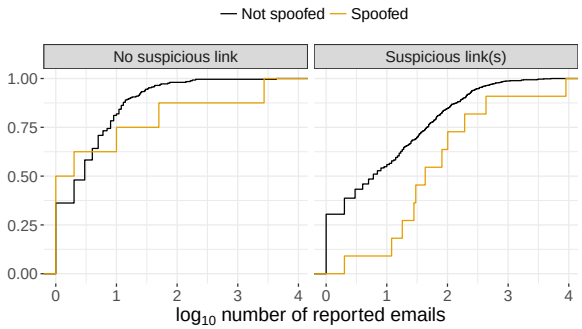


Figure 4.6: CDF of spoofed and non-spoofed `From:` domains

spoofed and non-spoofed `From:` domains for reported email with and without a suspicious link in the body. We observe a clear differentiation in the distribution, whereby emails with no suspicious link are approximately as likely to have a spoofed/non-spoofed `From:` address. A first finding is that emails with suspicious links are more likely to be delivered from non-spoofed than from spoofed addresses. This is compatible with a model of untargeted attacks where the attacker can be expected to be relatively unsophisticated. However, we find that a substantial amount of `From:` addresses are spoofed, with only about 25% being reported more than a hundred times. Overall, we find 12,701 unique spoofed `From:` addresses, making up for more than 17% of all observed `From:` addresses. This suggests that attackers (or one very dedicated attacker) spend considerable effort in generating new addresses (e.g. to avoid blacklisting) that closely resemble the target organization's domain.

4.3 Identification of campaigns

Figure 4.7 reports the distribution of suspicious emails that likely belong to the same campaign. Most campaigns are instantaneous in that they only last one day, with approximately 75% of similar emails arriving less than 10 days apart, and 90% of emails arriving less than 2 months apart with a relatively long tail. From the distribution it appears that *instantaneous* campaigns are common (up to one day), whereas unusually long campaigns extend for more than 100 days. Table 4.2 reports summary statistics of suspected phishing campaigns. Whereas most campaigns are instantaneous, we identify 46 distinct campaigns lasting on average 162 days (approx 5 months) and up to 264 days in the observation period.

While email bodies in `LONG` campaigns tend to be similar (i.e. introducing only limited text), `From:` addresses may evolve to a lower (on average) dissimilarity from the original domain. To

Table 4.2: Descriptive statistics of duration and intensity of phishing campaigns

Most phishing campaigns are instantaneous with only a fraction (8%) lasting more than 100 days. `INSTANTANEOUS` campaigns last up to one day; `SHORT` campaigns up to 100 days; `LONG` campaigns more than 100 days. The increasing number of samples for longer campaigns indicates that there is likely continuity in the delivered attacks.

		Phishing samples							Campaign duration (days)						
Type	n	Min	1stQ	Mean	Med	3rdQ	Max	sd	Min	1stQ	Mean	Med	3rdQ	Max	sd
`INST.`	373	1	1.0	12.0	1	2.0	622	52.3	0	0.0	0.1	0.0	0.0	1.0	0.2
`SHORT`	127	2	4.5	142.2	38	130.5	1394	266.0	1	6.6	31.4	22.0	48.5	98.2	27.8
`LONG`	46	2	65.0	878.6	227	983.5	6070	1386.8	101	118.8	162.1	150.8	186.4	264.9	44.6

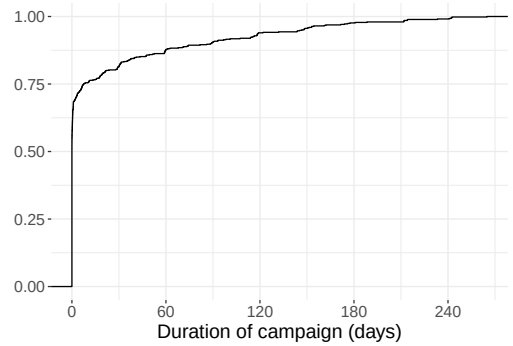


Figure 4.7: Duration of phishing campaigns

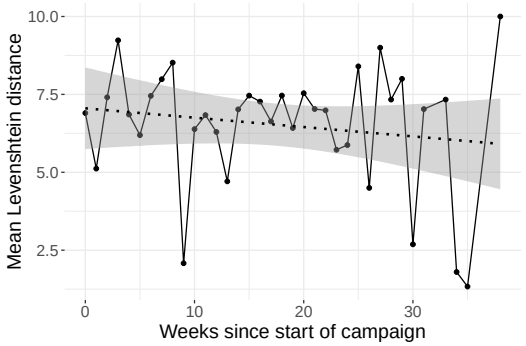


Figure 4.8: Decrease in dissimilarity between spoofed domain and Rabobank name

investigate how spoofing evolves during campaigns, Figure 4.8 reports the weekly average similarity between the domain of the attacker `From:` address and the domain of the victim organization (measured as their Levenshtein distance) for `LONG` campaigns. Lower scores are better. We observe an average decrease in dissimilarity between spoofed `From:` addresses and organization domain, which suggests an overall learning for attackers that in the long run may adjust or refine phishing attacks while retaining the same phishing content. Interestingly, we observe that in the long run appear sparse bursts of more-than-average sophisticated campaigns with well spoofed `From:` addresses ($cor = -0.23, p = 0.04$). Treating these episodes as outliers does not change the finding that spoofing dissimilarities decrease as campaigns advance ($cor = -0.20, p = 0.07$).

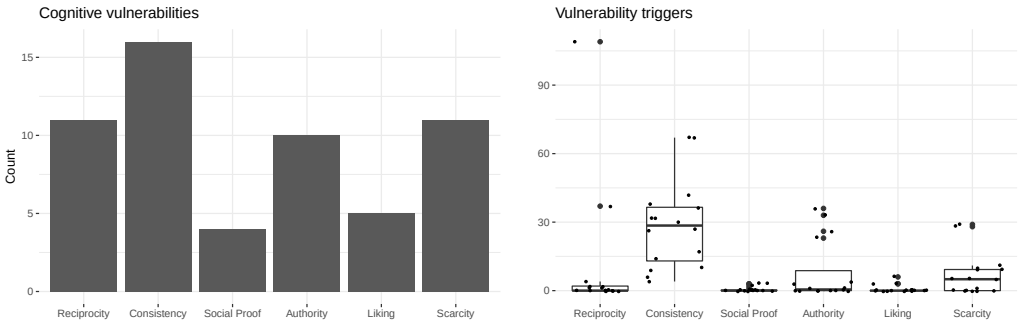
Chapter 5

Cognitive evaluation of phishing attacks

We report below an example of a phishing email (translated to English) and its association with different cognitive vulnerabilities. Vulnerability triggers are indicated in *italics* and refer to (1) `Liking`, (2) `Consistency`, (3) `Authority`, (4) `Social Proof`, (5) `Reciprocity` and (6) `Scarcity`:

> *(1) As a valued consumer of Rabobank* we always want to inform you of the latest updates and innovations in our system. We have recently switched to a new system that requires *(4) all current customers* to replace their *(2) current debit cards* by our newly-produced ones.
>
> In connection with the new changes to the *(3) European Safety Regulations*, Rabobank wishes to attend all its customers to the availability of the new and improved debit cards that adhere to all *(3) environmental and safety regulations.*
>
> *(1) Rabobank strives to be environmentally friendly.* Therefore, our service team will recycle all current debit cards by mounting your *(2) current AES Encryption Chip* on your renewed biological RFID payment card. For this reason, all current payment cards must be replaced. *(5) By participating in our recycling program, the new debit card can be requested free of charge.* *(6) After October 19th, 2018, a direct debit will be charged.*

From the example we can observe that the different cognitive vulnerabilities often appear alongside each other, and that a single vulnerability can even occur multiple times within an email body. Figure 5.1 reports the distribution of triggered cognitive vulnerabilities in each manually labeled email (left) and the corresponding vulnerability triggers identified in the corpus (right). We observe a clear relation between the two plots: the most common vulnerabilities and triggers in emails embedding successful links appear to be linked to the `Consistency` and `Scarcity` vulnerabilities. `Liking` and `Social Proof` triggers appear to be particularly rare, with most emails targeting none. This is consistent with the intuition that in one-shot interactions (as



Most attacks trigger `Scarcity`, `Consistency`, and `Reciprocity` vulnerabilities. `Social Proof` and `Liking` are the least common. Relative frequency of cognitive vulnerabilities is reflected in the distribution of vulnerability triggers identified in the emails.

Figure 5.1: Distribution of triggered cognitive vulnerabilities (left), and of vulnerability triggers (right)

opposed to prolonged or repeated exchanges as in spear-phishing attacks [4]) cognitive attacks linked to the target's social context and personal preferences (ref. Table 2.1) are rare as these are harder to implement with short interactions. By contrast, exploiting `Consistency` may only require reference to previous actions that the group of potential victims will have likely performed, such as buying an insurance or receiving a debit card from the organization. `Authority` appears to be a relatively common trigger in our sample, albeit not for all emails. Common triggers here refer to European and national-level legislation and often come together with the threat of a punishment if certain actions are not completed.

To evaluate the effect of the cognitive features of the email(s) embedding the URL links and the recorded clicks, we first report in Figure 5.2 the distribution of average clicks generated by emails. Most emails generate fewer than 150 clicks, with two emails generating more than 200 clicks ($min = 6,\ median = 78,\ max = 260,\ sd = 73.42$). A first finding is that, out of the $5,600$ emails obtained in the second sample (ref. Table 3.1), we find only twenty matches (or 0.36%) to a triggered external URL. This is in line with previous work reporting very low click-through rates for spam and phishing campaigns [74, 75].

Figure 5.3 displays the relation between triggered cognitive vulnerabilities and generated clicks, for which we observe a clear positive relation.[1] Following common practice [77], to avoid dispersion we only consider URLs clicked at least ten times, and remove four emails whose links only generated

[1] A possibility is that some emails may be distributed to substantially more users than others, generating different aggregate click counts. As we have no access to the victim's inboxes, we cannot directly measure this. However, the data does not show specific biases in the likelihood of users reporting emails (Fig. 4.5), suggesting that major skews are not realistic. Importantly, this is consistent with previous findings in the literature on (untargeted) phishing and malware attacks, whereby criminals relied on shared pools of targeted users [76, 41] as opposed to cherry-picking which users should receive which email. Further, due to the very low click-through rates of spam and phishing campaigns [74], this difference should be of several orders of magnitude between users to have a visible effect (as opposed to be undetectable noise in the data generation process).

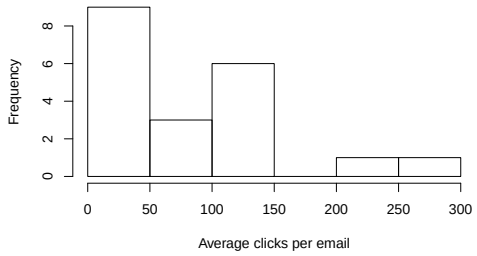
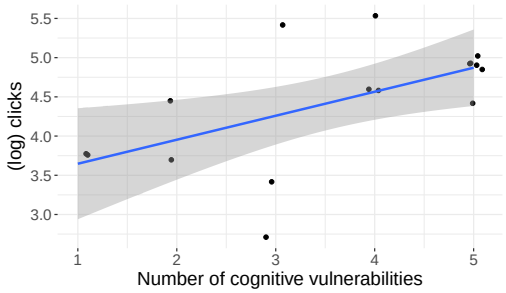
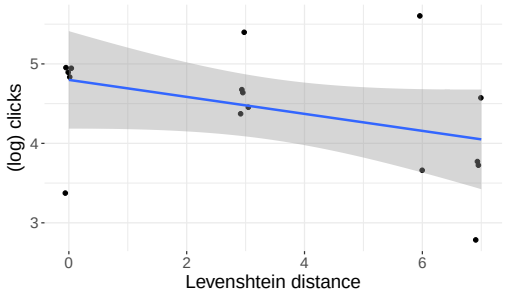


Figure 5.2: Histogram distribution of clicks per email



We observe a clear relation between the presence of exploited cognitive vulnerabilities and the clicks generated by the URLs embedded in the phishing emails.

Figure 5.3: Relation between number of vulnerabilities and average clicks (log_{10})

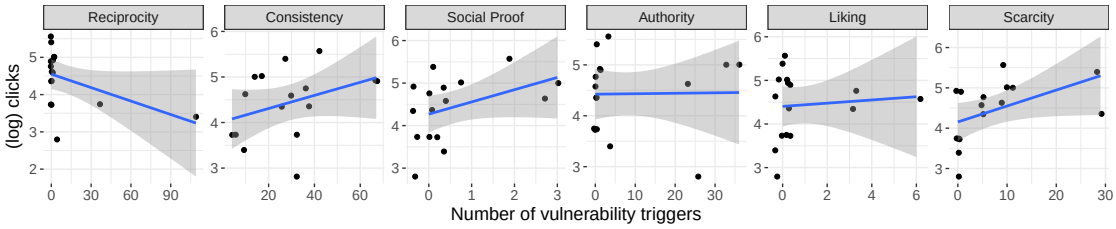


We identify a negative relation between the dissimilarity of the spoofed `From:` domain in an email against the original one, and the expected number of clicks.

Figure 5.4: Relation between spoofing dissimilarity and average clicks (log_{10})

six clicks. A simple Poisson regression of the form $log(clicks_i) = \alpha + \beta(cogvulns_i)$ reveals a strong correlation between the variables ($\beta = 0.25, p < 0.001$). This suggests that the more cognitive vulnerabilities are exploited in an email body the more that email can be expected to generate complying user behaviour from the intended target.

Apart from the cognitive vulnerabilities exploited in the text, a second relevant factor could be the similarity of the `From:` address displayed to a user and the legitimate one. The median Levenshtein distance between the spoofed domains and the original one in the successful emails is 3 (i.e. three substitutions in the spoofed domain are required to match the original domain). This is significantly smaller than for all emails in the corpora ($p = 0.001$, $median = 7$). Figure 5.4 reports the relation between Levenshtein distance of the spoofed `From:` domain and the expected number of clicks. We find an inverse relation between the two variables, suggesting that the greater the dissimilarity between the spoofed and the original domain, the lower the average number of generated clicks ($\beta = -0.06, p < 0.001$). This suggests that both cognitive attacks and the degree of spoofing in an email may have an effect on the relative success of a phishing email and could be considered to build a triaging model for phishing emails.



The data shows the effect of different cognitive vulnerability triggers on expected number of clicks. `Consistency`, `Social Proof`, and `Scarcity` have a clear positive association with the expected number of clicks they generate. `Authority` and `Liking` do not show any evident trend. Interestingly, we find that `Reciprocity` appears to be counter-productive, however we note that this effect is only driven by a few non-zero data points.

Figure 5.5: Correlation between vulnerability triggers and observed clicks

5.1 Analysis of vulnerability triggers

To evaluate the success of a phishing email we consider the relation between observed number of clicks and the presence of vulnerability triggers for each cognitive attack. Figure 5.5 reports the results. The data reports a clear positive relation between `Consistency`, `Social proof`, and `Scarcity` vulnerability triggers with the expected (log) number of clicks. `Reciprocity` shows a negative relationship, but that appears to be driven by few data points only whereby the evident majority of emails have relatively small counts of triggers for this cognitive vulnerability (see also Fig. 5.1). Likewise, both `Authority` and `Liking` suffer from a lack of meaningful data points which could explain the absence of a clearly observable effect.

5.2 Model selection

We now evaluate the relative impact of each coefficient in the collected dataset. We estimate coefficients for a Poisson process of the (aggregate) form:

$$log(clicks_i) = \alpha + \beta_1 cogvulns_i + \beta_2 spoofdist_i + \epsilon_i \tag{5.1}$$

whereby, for each email i, *clicks* represents the number of generated clicks, *cogvulns* is the array of counts of the vulnerability triggers identified in the email body, and $spoofdist$ indicates the degree of (dis-)similarity between the spoofed `From:` address and the original Rabobank domain. ϵ_i is the error term. To monitor and account for overfitting problems related to the few available datapoints, we combine a step analysis of each model (M1..M7) with regression bootstrapping to generate robust confidence intervals for the coefficient estimations.

Table 5.1: Regression results for Eq. 5.1

All model coefficients estimations are relatively stable across the seven models. Coefficients for the Poisson models are presented with 95% confidence intervals in parentheses. `Social proof` and `Spoof distance` `From:` addresses appear to have the largest effects on predicted number of clicks. Higher spoof distances (i.e. higher dissimilarity between `From:` domain and original domain) result in a lower number of expected clicks. Model power w.r.t. the baseline model is reported by the adjusted McFadden Pseudo-R^2 [79]. We only report significance coefficients (indicated by a * for significance at 0.1% level) for the reader's reference; however due to the relatively small sample size coefficient estimations should only be interpreted relative to each other as opposed to in absolute terms. Standard checks on the distribution of the residuals do not reveal issues or biases in the model fit.

	M1	M2	M3	M4	M5	M6	M7
α	4.75*	4.48*	4.08*	4.19*	3.60*	3.42*	4.24*
	(4.70, 4.80)	(4.39, 4.58)	(3.96, 4.21)	(4.06, 4.31)	(3.41, 3.78)	(3.18, 3.65)	(3.86, 4.60)
`Reciprocity`	-0.01*	-0.01*	-0.01*	-0.01*	-0.01	-0.01 .	-0.01*
	(-0.02, -0.01)	(-0.01, -0.01)	(-0.01 -0.01)	(-0.01, -0.01)	(-0.01, -0.01)	(-0.01, 0.01)	(-0.01, -0.01)
`Consistency`		0.01*	0.01*	0.01*	0.02*	0.02*	0.01*
		(0.01, 0.01)	(0.01, 0.01)	(0.01, 0.01)	(0.01, 0.02)	(0.01, 0.02)	(0.01, 0.01)
`Social proof`			0.27*	0.34*	0.33*	0.35*	0.44*
			(0.22, 0.31)	(0.28 0.40)	(0.28, 0.39)	(0.29, 0.41)	(0.37, 0.51)
`Authority`				-0.01*	-0.01 .	-0.01	-0.02*
				(-0.01, -0.01)	(-0.01, 0.01)	(-0.01, 0.01)	(-0.03, -0.01)
`Scarcity`					0.03*	0.03*	0.02*
					(0.02, 0.03)	(0.03, 0.04)	(0.02, 0.03)
`Liking`						0.05	0.01
						(0.01, 0.08)	(-0.02, 0.05)
`Spoof dist.`							-0.08*
							(-0.11, -0.05)
Adj. Pseudo-R^2	0.17	0.23	0.40	0.43	0.59	0.60	0.63
BIC	636	593	460	443	317	312	286

Table 5.2: Correlations between regression variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) `Reciprocity`	1.00	-0.25	-0.17	-0.10	-0.15	-0.27	-0.23
(2) `Consistency`		1.00	-0.27	-0.32	0.12	-0.11	-0.35
(3) `Social proof`			1.00	0.67	-0.24	0.15	0.04
(4) `Authority`				1.00	-0.28	0.04	-0.13
(5) `Liking`					1.00	-0.10	-0.07
(6) `Scarcity`						1.00	-0.13
(7) `Spoof dist.`							1.00

For model selection we report coefficients, 95% confidence intervals, BIC scores [78], and Adjusted McFadden Psuedo-R^2 [79], to reduce the statistical bias in the performance metrics for model selection.[2] Results are reported in Table 5.1.

All models have relatively stable coefficient estimations showing no evident interaction effect between the regressors as can be seen from the correlation matrix presented in Table 5.2. Coefficients should be interpreted relative to each other as opposed to in absolute terms. Because of the relatively small sample size, we refrain from drawing direct conclusions on the model coefficients. For this reason statistical significance is better served in the analysis reported in Figure 5.5 and is only detailed in Table 5.1 for the reader's reference. Within our sample, model coefficients can be interpreted as the relative change in number of clicks for every new vulnerability trigger. For

[2]Importantly, with this procedure we *do not* aim at identifying a definitive model and coefficients to forecast phishing success: regardless of the amount of observations in the dataset, that would not be possible because the 'click generation process' generating the observations necessarily varies from domain to domain (e.g. finance vs health), from organization to organization (e.g. national vs international), and from customer base to customer base (e.g. sensibility of application domain). Therefore, coefficient estimations out of this type of models cannot be '*plug-and-play*' across organizations and domains and will require tuning before being applied in-house.

example, the M7 coefficient for `Scarcity` (0.02) indicates an increase of 2% in the number of expected clicks for every new trigger of that category. Likewise, an increase in one point on the Levenshtein distance scale is related to a *decrease* in clicks of 8%. Looking at the BIC scores and at the McFadden's $Pseudo - R^2$s, `Social Proof`, `Scarcity`, and `Spoof distance` appear to have the strongest effect in increasing the explanatory power of the model. This is also reflected in the higher estimation for the variable coefficients. The negative effect of `Reciprocity` is confirmed in the model as well. `Authority` also appears to have a negative effect in all the models, suggesting that additional interactions with the context of the email may be relevant here for the credibility of the phish. In all, `Authority` and `Liking` appear to have the smallest effect on the model. However, with reference to Figure 5.5, the estimations of `Reciprocity`, `Authority` and `Liking` can not be considered accurate due to a lack of meaningful data points.

5.3 Cognitive triaging of phishing success

We now extend the model evaluation to estimate the amount of clicks generated by other emails for which Rabobank has detected no click (e.g. because no call-back to Rabobank resources originated from the phishing website, remaining therefore invisibleto Rabobank's detection infrastructure, ref. Fig 3.1). Recall however that our model estimates are likely subject to overfitting issues due to the inevitably small sample size. This only means that predicted outcomes could be unreliable over arbitrarily diverse email corpora (i.e. not represented in the training data); on the other hand, predictions over *similar* emails to those provided to the fitted model above will not suffer from unmodelled biases and will generate reliable estimations. For this reason we only limit our analysis to emails with a distribution of vulnerability triggers within plus or minus one standard deviation from the mean for that trigger in the model's respective training set.

Following the BIC model selection described above we find `Liking`, `Authority` and `Reciprocity` to be generally irrelevant in our scenario as we can observe from Figure 5.5 that most of our sample data points have 0 triggers recorded for these vulnerabilities. Based on this observation we consider four prediction models (PM), each with different regressors, namely (1) all three relevant cognitive vulnerabilities, (2) all three relevant cognitive vulnerabilities + spoofing distance, (3) all six cognitive vulnerabilities, (4) all six cognitive vulnerabilities + spoofing distance.

Table 5.3: Bootstrapped regression coefficients

PM1 provides confident and quite precise estimates for the positive effects of `Consistency`, `Social Proof` and `Scarcity`. We find no evident contribution of `Spoof dist.` to overall model quality, and as expected `Reciprocity`, `Authority` and `Liking` introduce more uncertainty in the coefficient estimations.

	PM1			**PM2**			**PM3**			**PM4**		
	0.025q	Median	0.975q	0.025q	Median	0.975q	0.025q	Median	0.975q	0.025q	Median	0.975q
α	2.78	3.41	3.87	2.83	3.57	6.02	1.07	3.53	3.85	-13.75	4.48	19.41
`Reciprocity`							-0.43	-0.00	0.13	-0.42	-0.01	0.47
`Consistency`	0.01	0.02	0.03	-0.02	0.02	0.04	-0.01	0.02	0.06	-0.20	0.01	0.28
`Social Proof`	0.03	0.31	0.64	-0.10	0.31	0.68	-0.58	0.36	2.24	-5.72	0.45	6.26
`Authority`							-0.06	-0.00	0.18	-1.09	-0.03	0.58
`Scarcity`	0.01	0.04	0.15	0.00	0.04	0.14	-0.01	0.04	0.20	-0.26	0.02	0.23
`Liking`							-0.23	0.05	0.35	-1.70	0.01	0.59
`Spoof dist.`				-0.32	-0.04	0.05				-1.71	-0.11	2.28
N		3,985			1,271			3,258			1,156	

More specifically:

$$\text{PM1} = \alpha + \beta_1 consistency_i + \beta_2 socialproof_i + \beta_3 scarcity_i + \epsilon_i \tag{5.2}$$

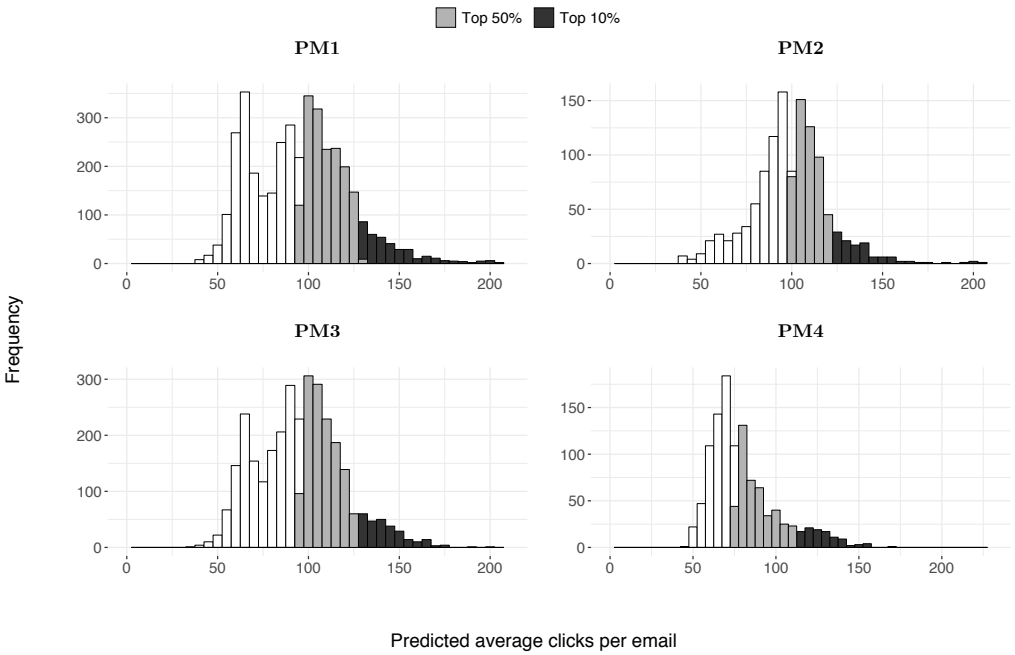
$$\text{PM2} = \alpha + \beta_1 consistency_i + \beta_2 socialproof_i + \beta_3 scarcity_i + \beta_4 spoofdist_i + \epsilon_i \tag{5.3}$$

$$\begin{aligned}\text{PM3} = \alpha &+ \beta_1 reciprocity_i + \beta_2 consistency_i + \beta_3 socialproof_i + \beta_4 authority_i + \beta_5 liking_i \\ &+ \beta_6 scarcity_i + \epsilon_i\end{aligned} \tag{5.4}$$

$$\begin{aligned}\text{PM4} = \alpha &+ \beta_1 reciprocity_i + \beta_2 consistency_i + \beta_3 socialproof_i + \beta_4 authority_i + \beta_5 liking_i \\ &+ \beta_6 scarcity_i + \beta_7 spoofdist_i + \epsilon_i\end{aligned} \tag{5.5}$$

We generate robust confidence intervals around the coefficient estimations using bootstrap simulations ($n = 10,000$). Table 5.3 reports median coefficients and 95% confidence intervals of the estimations for all four models. We compare PM1 with PM3 and PM2 with PM4 to explore the effect of keeping only the relevant cognitive vulnerabilities versus keeping all six cognitive vulnerabilities as regressors in the prediction models. PM1 provides us with specific and clearly positive confidence intervals for each of the relevant coefficients, whereas PM3 shows zero-overlapping and wider confidence intervals. Likewise, PM4 shows a significant reduction in preciseness of the estimations in comparison with PM2. These findings confirm our intuition that the exclusion of `Liking`, `Authority` and `Reciprocity` produces more robust estimations. However, note that coefficient estimations vary widely between runs, suggesting only minor estimation issues due to overfitting of the models.

Similarly we examine the effect of adding `Spoof dist.` as a regressor by comparing PM1 with PM2 and PM3 with PM4. From the first comparison, we find that PM2 provides slightly wider confidence intervals for the coefficients of `Consistency` and `Social Proof` than PM1. Additionally,



Although PM4 presents a slightly more skewed distribution of predictions, the predictions across the four models remain relatively stable, indicating confidence in the estimated ranges.

Figure 5.6: Distribution of predicted average clicks

the confidence interval for `Consistency` in PM2 is overlapping zero, indicating less confidence in the positive direction of the effect. Likewise, the addition of `Spoof dist.` to PM4 results in significantly wider confidence intervals in comparison with PM3. Therefore, we find that inclusion of `Spoof dist.` may lead to less confidence in the estimations due to the limited sample size.

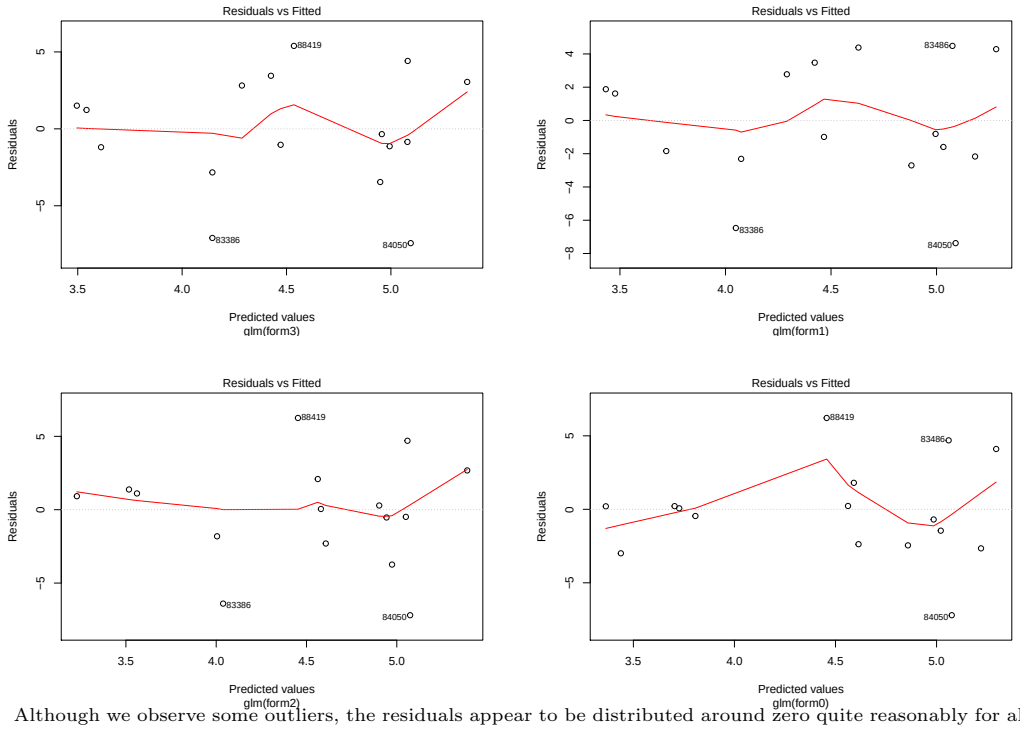
We simulate model predictions for the undetected clicks by randomly sampling ($n = 500,000$) model coefficients from the distributions of all four models and report aggregate statistics (Table 5.4) of the estimated number of generated clicks. Figure 5.6 visualizes these results. Although we observe that PM4 predictions are quite skewed in comparison with the other models that are more symmetric, the overall predictions across the four model variations remain relatively stable. Therefore we can be confident in the reliability of the estimated click ranges. The aggregated simulations indicate that the average 'undetected' email has potentially generated 80-100 clicks, with a long tail of (few) emails generating up to over 200 clicks. [3]

From these findings it is clear that, irrespective of the prediction model, a small portion of the attacks can be expected to be up to 2 times as effective as the bulk of incoming attacks. Therefore, prioritization efforts based on the cognitive characteristics of a phishing email could help in more

[3]Notice that additional organization-specific features of the email (e.g. presence of the company logo), may also have an effect on the number of clicks. Whereas this is out of the scope of this thesis, which only looks at the cognitive effects, a fully-operative model within an organization can easily integrate other factors in the prediction.

Table 5.4: Descriptive statistics of predicted average clicks

	Min.	0.025q	Median	Mean	0.975q	Max.
PM1	38	74	96	95	112	226
PM2	40	89	99	99	110	205
PM3	37	79	96	95	109	200
PM4	45	67	75	80	88	172



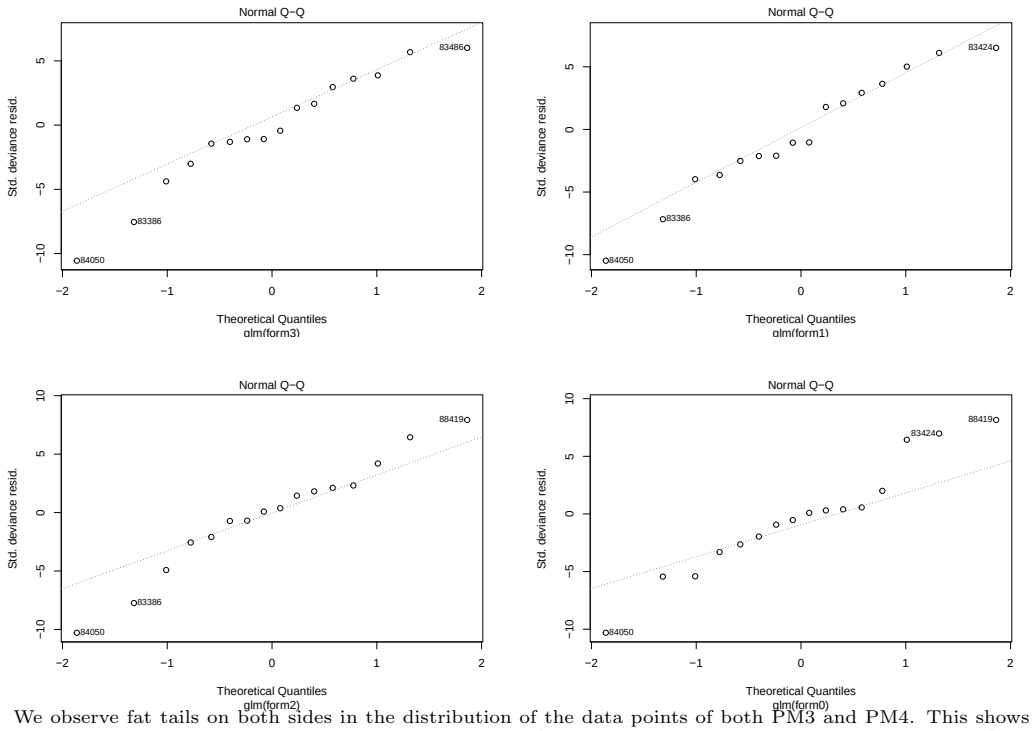
Although we observe some outliers, the residuals appear to be distributed around zero quite reasonably for all four models ($mean_{PM1} = -0.22$, $mean_{PM2} = -0.21$, $mean_{PM3} = -0.18$, $mean_{PM4} = -0.17$). Furthermore we find no evident patterns in the distribution of the residuals.

Figure 5.7: Residuals vs fitted values

efficiently addressing attacks (e.g. by means of takedown actions). By targeting first the emails that are most likely to engage users in compliant behaviour, organizations can effectively triage the stream of incoming phishing attacks to minimize the impact on their customer base.

5.4 Model checks

Figure 5.7 show residuals versus predicted values for each of the four prediction models. We observe no obvious patterns in either of the residual plots, although we can observe a slightly more uniformly distributed scatter of points in PM1 and PM2 in comparison to PM3 and PM4, which indicates more reliable predictions from these models. From the mean values of the residuals



We observe fat tails on both sides in the distribution of the data points of both PM3 and PM4. This shows that compared to an actual Poisson distribution, more of these data points are located at the extremes of the distribution as opposed to the center. PM1 and PM2 show more of a right skew, meaning that data points are closer on the left of the distribution and further away on the right.

Figure 5.8: Normal Quantile-Quantile plots

($mean_{PM1} = -0.22$, $mean_{PM2} = -0.21$, $mean_{PM3} = -0.18$, $mean_{PM4} = -0.17$) we find that addition of `Spoof dist.` in PM2 and PM4 seems to very faintly reduce residual error.

Figure 5.8 shows normal quantile-quantile (normal Q-Q) plots for each of the four prediction models. Note that normality of the residuals for Poissonian models can not generally be expected, so we refrain from drawing definitive conclusions from these plots. We observe fat tails in the distributions of both PM3 and PM4. Compared to a normal distribution, more of these data points are located at the extremes of the distribution as opposed to the center. PM1 and PM2 show more of a slight positive skew, meaning that data points are closer on the left of the distribution and further away on the right.

The outliers in the residuals and the findings from the normal Q-Q plots indicate some degree of overdispersion for each of the models, suggesting the Poisson distribution may be subject to small prediction errors. However, as we can characterize our data as consisting of independent observations of count data, a Poisson distribution is appropriate. Model runs assuming a quasi-poissonian fit lead to qualitatively and quantitatively similar results.

Chapter 6

Discussion

The previous sections have demonstrated how quantitative measurements of cognitive vulnerabilities employed in phishing attacks can be used to develop a model to make predictions about the expected efficacy of these attacks. This characterization allows one to assess the threat of these attacks in an automated way such that instant prioritization of phishing incident responses becomes possible. This thesis' contributions go beyond the scope of earlier works on cognitive factors for phishing by providing an empirical estimation and operable implementation of a triaging mechanism for prioritizing phishing incident response.

6.1 Discussion of exploratory results

In Chapter 4 we performed an exploratory analysis on the full collected email dataset. Interestingly, we found that the distribution times of phishing emails coincide with recommended sending times of email marketing newsletters. Therefore, not only do phishing attacks apply the same influence principles so commonly used for general marketing purposes, they also employ the same strategies used by online marketeers to maximize the reach of their email content. This suggests that unlike in the early days of phishing, when attacks could often be recognized by bad grammar and spelling (to narrowly target a lower-educated population [48]), more recent attacks have advanced to target more internet savvy audiences as well. This is corroborated by our finding that users are as likely to report emails that contain a suspicious link as emails without, which indicates that it may be exceptionally hard for users to distinguish between legitimate marketing emails and phishing emails. For example, phishing attacks may invoke a sense of urgency in a way similar to marketing

emails for limited-time product offers. Therefore, current phishing education and awareness may benefit from highlighting especially this overlap between general consumer marketing techniques and the social engineering efforts found in phishing attacks.

6.2 Discussion of cognitive evaluation results

In Chapter 5 we presented several correlations between different cognitive vulnerabilities and the average number of clicks an email can be expected to generate. We hypothesized that more extensive forms of impersonation and persuasion lead to more effective attacks. In line with this hypothesis we found that both higher degrees of impersonation in the phishing sender domains and the presence of any individual cognitive vulnerability increases user response to the phish. More specifically, we found that `Consistency`, `Social Proof`, and `Scarcity` exercise a clear positive effect on the number of generated clicks. We find no evident effect from `Authority` and `Liking`, and `Reciprocity` even shows a counterproductive effect, albeit only marginal, we note again however that these effects were driven by only few non-zero data points.

The overall findings contrast with previous studies on susceptibility to persuasion in phishing attacks, with one study [13] reporting `Liking` to have the largest impact on the likelihood of university students to respond to phishing attacks and another [14] showing `Authority` to most significantly impact phishing effectiveness. These difference may well be explained by the specific application domain, as corporate customers subject to financial threats from phishing can generally be expected to have different sensitivity to specific *principles of influence* than other groups [46]; the relative efficacy of the different cognitive vulnerabilities is context-dependent. Other demographics of the targeted individuals may play a role as well, with both age [50] and gender [50, 80, 81] being mentioned as important factors impacting an individual's susceptibility to persuasion in phishing attacks. Additionally, we can consider the finding that characteristics of the communicator are less evident in written communications [82], which could lead targets of influence to become more focused on the actual contents of the message than on the source of it. Therefore, influence triggers that rely on some degree of interpersonal interaction, may become less salient in phishing emails. Although this suggests that full generalizability can not be expected for any one set of results, although we suspect conclusions similar to ours could be drawn for specific contexts close in nature to the one in which Rabobank operates.

These observations also provide useful input to training campaigns regularly run by medium and large organizations in an attempt to increase their customers and employee's awareness of

the social engineering threat. On the one side, replications of this study in specific domains could reveal to which principles of influence the 'average' customer of an organization is more subject to; awareness campaigns run by the organization could then target those specific traits. For example, consumers particularly vulnerable to `Scarcity` could potentially benefit to know which are the organization's policies in terms of change deadlines and processes, such that an email stating unrealistic and short cutoff dates to react lose in credibility. On the other hand, the presented procedure could be applied both client and server side to automate the tagging of potential phishing emails for the enforcement of local or remote policies.

Furthermore, we have described how these observed effects can be used in the construction of a prediction model for the triaging of incoming phishing attacks. By enabling the triaging of incoming phishing attacks, our results will enable incident response teams to focus on the most prominent threats immediately, without having to manually filter out the noise from the bulk of irrelevant emails in their phishing abuse inbox, thereby minimizing reaction costs and increasing response effectiveness. The practicality of this is evidenced in Figure 5.6 where by addressing the top 10% of emails one mitigates hundreds of potential attacks per takedown action; by contrast, following a first-comes-first-served process (which is essentially random), one would most likely end up treating notifications in the mass of emails, substantially limiting the amount of prevented attacks per unit of effort.

6.3 Future work

The triaging mechanism presented in this thesis reveals plenty opportunities in terms of automated incident handling and security orchestration, e.g. by enabling incident handlers to apply automated follow up procedures to incoming phishing attacks that fall within a certain threat range. Reported measures on the vulnerability triggers that account for the threat level of a specific email can provide useful information on the handling of the email, which can serve as input for dynamic risk-based access control policies to limit immediate follow-up actions. Similarly, CSIRTs could implement automated network-level containment procedures based on the profile of incoming emails, and avoid additional (and unnecessary) victimization by delaying follow-up actions until the risk is cleared. Future work could focus on the exploration of such automated strategies.

Furthermore, the clear overlap between consumer marketing techniques and the influence tactics employed in phishing emails illustrate a need for new methods to enable users to swiftly assess the legitimacy of incoming emails. Organizations may find a solution in digital watermarking

mechanisms that allow users to check the integrity of emails based on the watermarking of resources that generally travel with email content such as logo images, as is suggested in [83]. Future work could focus on advancing such methods and the user training required, and assessing their effectiveness in reducing phishing victimization rates.

Evaluations of how the triaging mechanism proposed in this thesis can be integrated into Rabobank's current incident handling and response processes are currently ongoing.

6.4 Study limitations

In this research we have strived for scientific objectivity and comprehensive documentation of all research activities to ensure replicability. Regardless, we can identify several limitations of the work presented in this thesis.

Due to the nature of the collected email dataset and its extreme size, the absence of noisy 'junk' emails can not be guaranteed, despite best efforts to remove badly formatted emails, spam advertisements, and other irrelevant emails. We also consider the possibility of researcher bias, which may have been subconsciously introduced as a result of manual email labeling efforts performed by a single person in the absence of a second research proficient in Dutch.

Furthermore, we have stated before that due to infrastructural limitations the tracing of the phishing landing web pages from the URLs contained in phishing emails could not be performed in an automated way. On a best-effort basis, the tracing operation has been performed manually for three weeks, where each morning (close to 9AM UTC+1) all emails collected from the day before were processed. As phishing domains change domains often and swiftly, we were able to obtain only few matches between phishing domains and detected phishing emails to be used as samples in our cognitive evaluation. We compensate for this by means of the analysis methodology that explicitly accounts for the low incidence rate of URL matches

Chapter 7

Conclusions

In this thesis we presented an empirical method and evaluation of the effect of cognitive vulnerability triggers in phishing emails on the expected 'success' of an attack. We employed a unique dataset from Rabobank, a large financial organization in The Netherlands, fetching data from their phishing-response division. Our results indicate that response teams operations, such as take down actions against rogue phishing domains, could largely benefit from a (fully automated) cognitive assessment of the email body to predict relative success of the attack, given the relevant user base. Our findings and method could also be employed to deploy more effective training and awareness campaigns in response to the more prominent threats suffered by potential victims. Future work could explore automated response strategies to contain attacks and/or delay user response where needed.

Bibliography

[1] G. Stringhini and O. Thonnard, "That aint you: Blocking spearphishing through behavioral modelling," in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2015, pp. 78–97. 1, 3, 15

[2] G. Ho, A. S. M. Javed, V. Paxson, and D. Wagner, "Detecting credential spearphishing attacks in enterprise settings," in *Proceedings of the 26rd USENIX Security Symposium (USENIX Security17)*, 2017, pp. 469–485. 1, 3, 15

[3] S. Marchal, G. Armano, T. Gröndahl, K. Saari, N. Singh, and N. Asokan, "Off-the-hook: an efficient and usable client-side phishing prevention application," *IEEE Transactions on Computers*, vol. 66, no. 10, pp. 1717–1733, 2017. 1, 15

[4] S. Le Blond, A. Uritesc, C. Gilbert, Z. L. Chua, P. Saxena, and E. Kirda, "A look at targeted attacks through the lense of an ngo." in *USENIX Security Symposium*, 2014, pp. 543–558. 1, 29, 34

[5] R. Cialdini, *Influence: The Psychology of Persuasion*, 1984. 2, 11, 12, 13

[6] P. Glasziou and B. Haynes, "The paths from research to improved health outcomes," *Evidence Based Nursing*, vol. 18, no. April, pp. 36–38, 2005. 2, 11

[7] K. T. Trotman, A. M. Wright, and S. Wright, "Auditor negotiations: An examination of the efficacy of intervention methods," *Accounting Review*, vol. 80, no. 1, pp. 349–367, 2005. 2, 11

[8] R. Cialdini and N. Goldstein, "The science and practise of persuasion," *The Cornell Hotel and Restaurant Administration Quarterly*, vol. 43, no. 2, pp. 40–50, 2002. 2, 11

[9] R. E. Petty and J. T. Cacioppo, "The elaboration likelihood model of persuasion," in *Communication and persuasion*. Springer, 1986, pp. 1–24. 2

[10] N. Akbar, "Analysing Persuasion Principles in Phishing Emails," Ph.D. dissertation, 2014. 3, 12

[11] A. Ferreira, L. Coventry, and G. Lenzini, *Principles of persuasion in social engineering and their use in phishing*, 2015, vol. 9190. 3, 12

[12] M. Workman, "Wisecrackers: A Theory-Grounded Investigation of Phishing and Pretext Social Engineering Threats to Information Security," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 4, pp. 1–12, 2008. 3, 13

[13] R. T. Wright, M. L. Jensen, J. B. Thatcher, M. Dinger, and K. Marett, "Influence techniques in phishing attacks: An examination of vulnerability and resistance," *Information Systems Research*, vol. 25, no. 2, pp. 385–400, 2014. 3, 10, 14, 44

[14] M. Butavicius, K. Parsons, M. Pattinson, and A. McCormac, "Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails," in *Australasian Conference on Information Systems*, no. Hong 2012, 2015, pp. 1–11. 3, 14, 44

[15] D. Birk, S. Gajek, F. Gröbert, and A. R. Sadeghi, "Phishing phishers - Observing and tracing organized cybercrime," in *Proceedings of the 2nd International Conference on Internet Monitoring and Protection*, no. Icimp, 2007. 7

[16] M. Jakobsson and S. Myers, *Phising and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*, 2007. 7, 14

[17] E. Lastdrager, "From fishing to phishing," Ph.D. dissertation, 2018. 7, 8

[18] C. Herley, "Why do Nigerian Scammers Say They are from Nigeria?" *Weis*, p. 321, 2012. 7

[19] M. Jakobsson, "Two-factor inauthentication the rise in SMS phishing attacks," *Computer Fraud and Security*, vol. 2018, no. 6, pp. 6–8, 2018. 7

[20] S. E. Griffin and C. C. Rackley, "Vishing," in *Proceedings of the 5th Annual Conference on Information Security Curriculum Development*, 2008, pp. 33–35. 7

[21] A. Vishwanath, "Habitual facebook use and its impact on getting deceived on social media," *Journal of Computer-Mediated Communication*, vol. 20, no. 1, pp. 83–98, 2015. 7

[22] K. D. Mitnick and W. L. Simon, "The Art of Deception: Controlling the Human Element in Security," *BMJ: British Medical Journal*, p. 368, 2003. 8, 11

[23] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 4, pp. 2070–2090, 2013. 8

[24] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," *Communications of the ACM*, vol. 50, no. 10, pp. 94–100, 2007. 8

[25] J. S. Downs, M. B. Holbrook, and L. F. Cranor, "Behavioral response to phishing risk," in *Proceedings of the Anti-Phishing Working Group's 2nd Annual eCrime Researchers Summit*, 2007, pp. 37–44. 8

[26] E. M. Hutchins, M. J. Cloppert, and R. M. Amin, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains," in *Proceedings of the 6th Annual International Conference on Information Warfare and Security*, no. July 2005, 2011, pp. 1–14. 8

[27] D. Mundie, "Unintentional Insider Threat and Social Engineering," Tech. Rep., 2014. 8

[28] J. Klensin, "RFC5321: Simple Mail Transfer Protocol," Tech. Rep., 2008. 9

[29] S. Kitterman, "RFC7208: Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1," Tech. Rep., 2014. 9

[30] D. Crocker, Brandenburg InternetWorking, E. T. Hansen, A. Laboratories, E. M. Kucherawy, and Cloudmark, "RFC6376: DomainKeys Identified Mail (DKIM) Signatures," Tech. Rep., 2011. 9

[31] M. Kucherawy, E. Zwicky, and Yahoo!, "RFC7489: Domain-based Message Authentication, Reporting, and Conformance (DMARC)," Tech. Rep., 2015. 9

[32] M. Vasek, J. Wadleigh, and T. Moore, "Hacking is not random : a case-control study of webserver compromise risk," *IEEE Transactions on Dependable and Secure Computing*, vol. 13, no. 2, pp. 206—-219, 2015. 9

[33] R. Verhoef, "A Phisher's View of Phishing: U-Admin 2.7 Phishing Control Panel," 2018. 10

[34] T. Moore, R. Clayton, and R. Anderson, "The Economics of Online Crime," *Journal of Economic Perspectives*, vol. 23, no. 3, pp. 3–20, 2009. 10

[35] K. Thomas, D. Y. Huang, D. Wang, E. Bursztein, C. Grier, T. J. Holt, C. Kruegel, D. McCoy, S. Savage, and G. Vigna, "Framing Dependencies Introduced by Underground Commoditization," in *Workshop on the Economics of Information Security*, 2015, pp. 1–24. 10

[36] J. F. George, J. R. Carlson, and J. S. Valacich, "Media Selection as a Strategic Component of Communication," *MIS Quarterly*, vol. 37, no. 4, 2013. 10

[37] K. E. Stanovich and R. F. West, "Individual differences in reasoning: Implications for the ratinality debates?" *Behavioral and Brain Sciences*, vol. 23, pp. 645–726, 2000. 10, 11

[38] Anti-Phishing Working Group, "Global Phishing Survey 1H2012," Tech. Rep. October, 2012. 10

[39] R. T. Wright and K. Marett, "The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived," *Journal of Management Information Systems*, vol. 27, no. 1, pp. 273–303, 2010. 10

[40] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '06*, no. November 2005, p. 581, 2006. 10, 14

[41] PhishLabs, "Phishing Trends & Intelligence Report: Hacking the Human," Tech. Rep., 2018. 11, 29, 34

[42] A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Utility, Probability, and Human Decision Making*, vol. 185, no. 4157, pp. 141–162, 1975. 11

[43] J. M. Darley and B. Latané, "Bystander Intervention in Emergencies: Diffusion of responsibility," *Journal of Personality and Social Psychology*, vol. 8, no. 4, pp. 377–383, 1968. 12

[44] S. Milgram, *Obedience to Authority: An Experimental View*, 1974. 12, 14

[45] B. J. Sagarin and K. D. Mitnick, "The Path of Least Resistance," in *Six Degrees Of Social Influence: Science, Application, and the Psychology of Robert Cialdini*, 2012, ch. 3. 12

[46] P. Lawson, O. Zielinska, C. Pearson, and C. B. Mayhorn, "Interaction of personality and persuasion tactics in email phishing attacks," in *Proceedings of the Human Factors and Ergonomics Society*, vol. 2017-Octob, 2017, pp. 1331–1333. 14, 44

[47] T. Blass, "The milgram paradigm after 35 years: Some things we now know about obedience to authority," *Journal of Applied Social Psychology*, vol. 29, no. 5, pp. 955–978, 1999. 14

[48] M. Wu, R. C. Miller, and S. L. Garfinkel, "Do security toolbars actually prevent phishing attacks?" in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 601–610. 14, 43

[49] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham, "School of phish: a real-world evaluation of anti-phishing training," in *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 2009, p. 3. 14

[50] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 373–382. 15, 44

[51] S. Marchal and N. Asokan, "On designing and evaluating phishing webpage detection techniques for the real world," in *11th {USENIX} Workshop on Cyber Security Experimentation and Test ({CSET} 18)*. USENIX Association, 2018. 15

[52] R. Bhakta and I. G. Harris, "Semantic analysis of dialogs to detect social engineering attacks," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing*, 2015, pp. 424–427. 15

[53] Y. Sawa, R. Bhakta, I. G. Harris, and C. Hadnagy, "Detection of Social Engineering Attacks Through Natural Language Processing of Conversations," in *2016 IEEE Tenth International Conference on Semantic Computing*, 2016, pp. 262–265. 15

[54] K. Ding, N. Pantic, Y. Lu, S. Manna, and M. I. Husain, "Towards building a word similarity dictionary for personality bias classification of phishing email contents," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, IEEE ICSC 2015*, 2015, pp. 252–259. 15

[55] Wikipedia Contributors, "Lijst van Nederlandse banken." 21

[56] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc. New York, 1986. 22

[57] Z. S. Harris, "Distributional Structure," *WORD*, vol. 10, no. 2-3, pp. 146–162, 1954. 22

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2012. 22

[59] A. Singhal, C. Buckley, and M. Mitra, "Pivoted Document Length Normalization," in *ACM SIGIR Forum*, 2011, pp. 21–29. 23

[60] B. J. Wilson and A. M. J. Schakel, "Controlled Experiments for Word Embeddings," *CoRR*, 2015. 23

[61] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, 2008. 23

[62] B. Efron and R. J. Tibshirani, *Introduction to the Bootstrap*, 1993. 23

[63] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processin*, vol. 1, no. 6, 2009, p. 248. 25

[64] F. Mosteller and J. Tukey, "Data analysis, including statistics." in *Handbook of Social Psychology*, 1968. 25

[65] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine Learning*, vol. 88, no. 1-2, pp. 157–208, 2012. 25

[66] J. Fürnkranz, K. Brinker, E. L. Mencía, and E. Hüllermeier, "Multilabel Classification via Calibrated Label Ranking," *Machine Learning*, pp. 1–23, 2008. 25

[67] S. L. Blond, A. Uritesc, C. Gilbert, Z. L. Chua, P. Saxena, and E. Kirda, "A Look at Targeted Attacks Through the Lense of an NGO," in *Proceedings of the 23rd USENIX Security Symposium*, 2014, pp. 543–558. 27

[68] Mailchimp, "Insights from Mailchimp's Send Time Optimization System," 2014. [Online]. Available: https://mailchimp.com/resources/insights-from-mailchimps-send-time-optimization-system/ 28

[69] Campaign Monitor, "What Our Data Told Us about the Best Time to Send Email Campaigns," 2014. [Online]. Available: https://www.campaignmonitor.com/blog/email-marketing/2014/08/best-time-to-send-email-campaigns-by-device/ 28

[70] SendInBlue, "Best Time to Send an Email: User Data Study by Industry," 2017. [Online]. Available: https://www.sendinblue.com/blog/best-time-to-send-email/ 28

[71] Propeller, "The 2017 Email Marketing Field Guide: The Best Times and Days to Send Your Message and Get It Read," 2017. [Online]. Available: https://www.propellercrm.com/blog/2017-email-marketing-field-guide 28

[72] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966. 29

[73] J. Szurdi, B. Kocso, G. Cseh, M. Felegyhazi, and C. Kanich, "The Long Taile of Typosquatting Domain Names," in *Proceedings of the 23rd USENIX Security Symposium*, 2014, pp. 191–206. 29

[74] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage, "Spamalytics: an empirical analysis of spam marketing conversion," in *Proc. of CCS'08*, ser. CCS '08. ACM, 2008, pp. 3–14. 34

[75] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru, "Phi.sh/$ocial: The phishing landscape through short urls," in *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, ser. CEAS '11. New York, NY, USA: ACM, 2011, pp. 92–101. [Online]. Available: http://doi.acm.org/10.1145/2030376.2030387 34

[76] M. Yip, N. Shadbolt, and C. Webber, "Why forums? an empirical analysis into the facilitating factors of carding forums," 2013. 34

[77] J. F. Lawless, "Regression methods for poisson process data," *Journal of the American Statistical Association*, vol. 82, no. 399, pp. 808–815, 1987. 34

[78] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978. 37

[79] D. McFadden, "Conditional logit analysis of qualitative choice behaviour," in *Frontiers in Econometrics*, 1974, pp. 105–142. 37

[80] B. M. Okdie, R. E. Guadagno, P. K. Petrova, and W. B. Shreves, "Social Influence Online," *International Journal of Interactive Communication Systems and Technologies*, vol. 3, no. 1, pp. 20–31, 2013. 44

[81] R. E. Guadagno, N. L. Muscanell, L. M. Rice, and N. Roberts, "Social influence online: The impact of social validation and likability on compliance." *Psychology of Popular Media Culture*, vol. 2, no. 1, pp. 51–60, 2013. 44

[82] S. Chaiken, "Communication modality as a determinant of persuasion," *Journal of Personality and Social Psychology*, vol. 45, no. 2, pp. 241–256, 1983. 44

[83] M. Topkara, A. Kamra, M. Attallah, and C. Nita-Rotaru, "ViWiD: Visible Watermarking Based Defense Against Phishing," in *4th International Workshop on Digital Watermarking*, 2005, pp. 470–483. 46