# End User Reporting of Suspicious E-mails Using X-ARF

Master Thesis

**Adrian Metzner**

May 2017

Supervisors:

**Prof. Dr. Dieter Gollmann**

**Sven Übelacker**

Hamburg University of Technology

**Security in Distributed Applications**

`https://www.sva.tuhh.de/`

Am Schwarzenberg-Campus 3

21073 Hamburg

Germany

# Declaration

I, Adrian Metzner, solemnly declare that I have written this master thesis independently, and that I have not made use of any aid other than those acknowledged in this master thesis. Neither this master thesis, nor any other similar work, has been previously submitted to any examination board.

Hamburg, May 7, 2017

Adrian Metzner

# Abstract

End user e-mail abuse reporting rates are very low, however a survey conducted in this thesis revealed that the desire to report abusive e-mails exists. The reason why the report rates are so low is that the process to report abusive e-mails is very time consuming and complex. Therefore, a new process for e-mail abuse reporting is defined. In this new reporting process, end users are able to report suspicious e-mails instead of spam or phishing e-mails. Furthermore, abuse reports are automatically generated in the X-ARF format. But since automatically generated reports can contain sensitive information, the end user has to have the option of anonymizing reports before sending them. As a proof of concept for the feasibility of this reporting process a Thunderbird plugin was developed.
The X-ARF format of the report enables a very simple implementation of automation for the handling of these reports. Therefore, as another conceptual proof, a report aggregator was developed. Which is able, to perform basic validation functionality.

# Preface

This thesis is written as the completion of my masters education in computer science.
In the spring of 2016 my supervisor, Sven Übelacker, brought the abuse reporting topic to my attention. Doing the initial research on the topic, I noticed that the thematic of e-mail abuse reporting fascinated me and that beneficial research could be performed there. Thus me and my supervisor came up with the idea to enable end users to report abusive e-mails.
In the curse of this thesis I gained insight into the world of incident management and abuse reporting as well as a variety of e-mail abuse cases. This insight into all of these field further cemented my believe, that end user e-mail abuse reports would be extremely beneficial for the security of organizations as well as private individuals and might even be a step towards solving the spam and phishing problematic. Therefore, I hope that this thesis is able to provide insight into the problems end users face in abuse reporting and the benefits more reports from end users could provide.
Adrian Metzner

# Contents

# 1 Introduction

The history of e-mails began in August 1982, when Jonathan B. Postel [55] proposed the *Simple Mail Transfer Protocol* (SMTP). This standard could be used to send text messages through the Arpanet, a precursor of the Internet. In 1996, the *Multipurpose Internet Mail Extensions* (MIME) standard was added. This new addition allowed for the sending of e-mails in different encodings, which led to a worldwide adoption of e-mail communication in private as well as professional communication. Today e-mail has become one of the most important communication media in our private and professional lifes.

But the success of the e-mail made criminals aware of this communication medium as well. At first, they started to abuse e-mail communication by sending masses of unsolicited advertisement e-mails. This became so bad that in 1997 Jill Lesser, a lawyer of AOL, reported at the FTC spam hearing that the rate of spam messages at most days during spring and early summer were around 30% [18] of all e-mails. Another form of abusive e-mails soon followed. Instead of simply sending unsolicited advertising, criminals focused on sending deceptive e-mails, which deceive the recipient into performing an action or giving away information. These e-mails are called phishing e-mails. While the damage of spam e-mails for an individual is rather low, phishing e-mails have the potential to cause huge amounts of damage.

Nevertheless, e-mail users do not remain defenseless against these threats. E-mail providers and researchers are constantly developing new ways and perfecting old ways to protect users from spam and phishing e-mails. One of the most successful defensive measures are the automatic spam filters. These filters try to categorize e-mails automatically into one of two categories, legitimate e-mails and spam or phishing e-mails. These filters operate by comparing new e-mails against already categorized ones and decide according to the highest similarities the filter can find. This has proven to work very well and filter most of the spam and a large number of phishing mails from the user's in-box. However, phishing and spam e-mails are always changing in order to bypass theses filters. Therefore, the filters have to update their training sets constantly.

Up to date training sets can be created through crowd sourcing. This approach relies on the users to report spam and phishing e-mails, which they received in their in-box, as such. Such reports are sent to trusted authorities, for example the mail provider of the end user, which can then use the information to train spam filters. Furthermore, other countermeasures against phishing and spam, such as blacklists, are also maintained through the reports. Last but not least, these reports do also allow for the take-down of phishing websites and closing of spammer accounts. This should in theory suffice to reduce the spam and phishing problem drastically. However, this is not the case.

One reason why the defenses can not employ their full effect is that not enough abusive e-mails are

reported. Most public e-mail providers, for example Google and GMX, allow their users to report spam e-mails with the simple click of a button. This feature is used quite frequently for advertisement spam such as the large Viagra campaigns of the past. Therefore, their spam filters are very good at filtering such e-mails. However phishing e-mails are rarely reported. One public clearinghouse, which accepts phishing reports is phishtank [53]. This publicly available clearinghouse has only about 90000 users, from which ten individuals have reported over 75% of the phishing sites and e-mails. This small number of reporters creates a huge problem, because most spam and phishing e-mails will never reach these individuals and therefore will not be reported. To increase the number of reported spam and phishing mails it is thus important to enable and incentivize end users to report abusive e-mails as well.
The focus of this thesis is to determine why end users do not report abusive e-mails and how they can be encouraged or enabled to report those e-mails. Therefore, a short survey was conducted. In this survey, it was tested how well the end users were able to detect phishing e-mails. Furthermore, the participants were questioned about their knowledge of incident reporting and their willingness to report abusive e-mails. This information is then used to describe a new e-mail abuse reporting process. It is specifically designed to enable and encourage end users to report abusive e-mails. Additionally as a proof of concept for the reporting process, a reporting tool for the popular Thunderbird mail client was developed, as an Add-on.
It is not part of the thesis to conduct usability studies regarding the Add-on or the reporting process. The reason for this is the limited time frame of six months in which this thesis was created. Furthermore, in this thesis it is also not researched how the introduction of the proposed process or the Add-on changes reporting behavior of end users over a longer time frame.

## 1.1 Overview

The thesis is structured as follows. The Chapter 2 provides an introduction into the context of abusive e-mails as well as incident reporting. Therefore, in Section 2.1 a short overview about abusive e-mails is given as well as the methods used to fight the different types of e-mail abuse. In the then following Section 2.2 the basics of incident reporting as well as reporting formats will be described. In Chapter 3 the conducted survey as well as its results are presented. The survey focused on the ability of end users to detect phishing e-mail as well as their desire to report abusive e-mails. Chapter 4 discusses the problems end users face when reporting abusive e-mails. Moreover, in Section 4.3 a new reporting process is presented, which takes the needs of end users into account and enables them to report abusive e-mails. Chapter 5 then evaluates this new process by comparing it to the old process. Last but not least, Chapter 6 summarizes the previous chapters and provides an outlook on future work that has to be done.

# 2 Preliminaries

## 2.1 Abusive E-mails

E-mails have become a widely used form of communication both in the business world as well as our private lifes. The reason for this has to be the simplicity of communicating via e-mails as well as the ability to contact multiple people at once without having to know where they live or who they personally are. However, these features do not only attract legitimate users but also individuals who seek to use the anonymity and mass communication capabilities of this media for their personal advantage. Thus, soon after the creation of the e-mail communication media some users abused the system.
Abusing the system is in this thesis defined as any e-mail message, which was received unsolicited and does not serve as a meaningful source of legitimate information or communication for the receiver. Thus examples for such e-mails are typical spam advertisements as well as phishing e-mails. Nevertheless, also threat or harassment e-mails count as abusive e-mails.
In this chapter, different types of abusive e-mails are discussed. Furthermore, the available countermeasures against these abusive e-mails are presented.

### 2.1.1 Spam E-mails

Spam e-mails are in most scientific papers defined as any unsolicited e-mail [81, 11]. However, this can be misleading, as in some cases an e-mail might be received unsolicited but can still be legitimate. An example would be the reception of a company newsletter, which is send to all employees. Therefore, in this thesis an e-mail has to be received unsolicited and not serve as a meaningful source of legitimate information or communication for the receiver. However, according to both definitions, all abusive e-mails can be defined as spam e-mails. Therefore, every countermeasure against spam e-mails will also work against all other subtypes of abusive e-mails. Subtypes of spam are for example phishing and harassment e-mails. However, harassment as well as threat e-mails behave just like all spam e-mails with the added twist, that additionaly harassment or threat laws apply to such e-mails [27]. Therefore, in this thesis there will be no distinction between these two and similar subtypes of spam and the overall term spam. Phishing e-mails however will be discussed in detail in the Section 2.1.3. In this thesis, the term “spammer” refers to individuals and organizations, which either directly send unsolicited e-mails or control any kind of automated tool, such as spam sending servers or botnets sending such e-mails.
The word spam is thought to originate from a sketch performed by the British comedy group *Monty Python's Flying Circus* in which a waitress reads a couple of customers the menu in which every item contains more and more spam, while a group of vikings in the background start singing “spam, spam, spam . . . ” louder and louder until the waitress can not be understood anymore [32]. This relates to the nature of spam e-mails, which without filtering reach such a volume that a victim can not manually

sort through his or her e-mails and identify the important solicited e-mails he or she receives.
To understand how spam e-mails work and why they exist, it is useful to study the history of spam. Even though we commonly think of spam as solely a feature of e-mail communication, the first steps towards spam started in chat programs. In those chats users could post arbitrary text in such quantities, that the legitimate communication disappeared from the screen and the rather weak network connections become exhausted [11].
Then came the usenet, which was a series of discussion channels in which interested people could read and discuss news or other topics they were interested in. But those usenet discussion forums allowed advertisers to post their advertisements into many different channels with very little effort and reach a mass audience. The most infamous case of such a posting is most likely the Canter and Siegel spam from 1994 in which the American immigration law firm *Canter & Siegel* [12, 18] used a computer program to post an advertisement for their services into 5.500 newsgroups in a very short time [12]. This lead the usenet community to filter their channels and automatically detect and report post that were made to multiple channels.
However, advertisers soon noticed the emerging e-mail communication medium and started sending advertisement via e-mails. Since e-mail communication is based on the communication between different computers, mails can be generated and send completely automatically. Therefore, the trend for automated spam sending, which started in the usenet forums, continued with spam e-mails. Spam senders (spammer) were able to send massive amounts of e-mails in a very short time, leading to a huge increase in unsolicited communication. In early summer of 1997 AOL reported, that about 30% of its traffic was caused by spam e-mails [18].
These volumes of spam can only be reached due to the extremely low cost of sending such an e-mail. Modern technology and software allows a single spammer to sent multiple thousands of e-mails in an hour. Thereby reducing the cost per e-mail that is sent below 0.0001$. The highest expense spammers have to pay is for obtaining valid e-mail addresses. But still a few thousand e-mail addresses have a price tag of only a few hundred dollars and can be reused in every spamming campaign [18].
Even though the cost for sending spam e-mails is quite low, the cost for the receiving side is a lot higher. The damages for the recipents can be calculated by first adding up the time individuals need to sort through their e-mail in-box and filter for such spam messages. Then the price for installing and maintaining automated spam filters, which are mandatory in order to keep e-mail communication viable with the current spam volumes, has to be included as well. In the paper *The Economics of Spam* [58] the costs for the spam receiving American companies and customers alone is estimated to be 20 billion dollars annually. However, world wide, the costs spam generates are estimated to be about 50 billion dollars annually [3, 2].
In the paper *Economics of Spam*, written by Rao and Reiley, the economic aspects of spam were analyzed. In the paper, it is estimated that spam only generates about 200 million dollar in profit each year. With the costs spam produces and the profits it generates, Rao and Reiley calculate the externality ratio of spam. Externality ratios are defined in economics as the relation between the costs of an action that are generated for those not actively involved and the benefits generated by that action. An example for

this would be an industrial complex which generates pollution. This complex creates profits, which are the benefits of having this complex. On the other hand the complex creates pollution, which affects the health and well being of people. The externality ratio now is the damage created through pollution divided by the profit of said complex. The smaller this ratio becomes the more socially beneficial is having this complex. Another example from the paper *Economics of Spam* is the externality of driving a car. The cost for driving one mile in a car are estimated as 0.25$ while the gain of this activity is given as 0.6$, establishing a cost to benefit ratio of around 0.5. Furthermore, stealing a car in the USA is in the same paper estimated to have a cost to benefit ratio between 6.70 and 30.3. Using these calculation on spam e-mails with the low cost estimate of 20 billion dollars annually, results in an externality ratio of 100.

### 2.1.2 Fighting Spam

The extremely high externality ratio shows the importance of fighting spam successfully. The success of fighting spam is defined in this paper defined as reducing the amount of spam the average e-mail account holder has to sort through. The techniques used to fight spam are always evolving alongside the techniques used by the spammers themselves.

When spam messages started to become a problem, e-mail providers started to notice that these messages were sent through their services anonymously. This was possible since sending an e-mail did not require any authentication [58], which was only required when reading e-mails. Therefore, spammers and other abusers could easily impersonate trusted companies and persons by using the same e-mail server as the impersonated individual or organization. Furthermore, since no authentication was necessary, spammers who do not want to impersonate anyone could use any mail server they liked to send their spam. Thus the first step to fight spam was to authenticate users, who wanted to send e-mails.

This did not stop the spammers for long, as they simply started to set up their own e-mail servers to circumvent authentication. Thus these spam e-mails were not sent from large trusted e-mail servers anymore, but rather from many small short lived mail servers. In order to handle the new wave of spam e-mails from those servers, filtering mechanisms were introduced. These first algorithms were largely based on crowd-sourcing and IP blacklisting. Therefore, only entire servers or single addresses could be blocked as spam sources. This allowed spammers who used servers which send mostly legitimate e-mails to bypass the new filters with ease. But in the 1990s maschine-learning based filters were added to the filtering algorithms [62] .

These filters allow for a fast and adaptive filtering process that is able to filter messages based on content and not solely by their source. These more complex filters were necessary since spammers, noticing the inefficiency of their personal e-mail servers, switched to new methods to distribute their spam. One of these new strategies is to automatically create accounts with freemail providers, such as Gmail or Yahoo, and use those addresses to send spam e-mails. This provides the spam e-mails with the trusted freemail provider's servers as the sending server. Thus filtering for the origin servers was not possible and the originating addresses can be easily switched, rendering the old filtering techniques useless.

However, the new machine-learning filters are able to sort through the e-mails and identify likely spam e-mails by matching them against certain patterns in the e-mail. The patterns used by these filters are generated through a subset of e-mails, which have been manually sorted into two clusters, one for spam e-mails and one for legitimate e-mails. These sample e-mails are then fed into the machine-learning algorithm, which derives rules from the already clustered sample to sort new e-mails into either the spam cluster or the legitimate cluster. The false positive and false negative rate of such filters depends on the samples they have got at to derive the rules from. Therefore, the best results can be achieved if the spam and legitimate e-mail sample sets are as diverse as possible regarding the content of the e-mails. An easy way to provide good sample sets is to use as many e-mails as possible, which generally increases the likelihood of diverse sample sets and reduces the sample bias that can occur when selecting a subset of samples from a larger set [45].

In order to obtain a set of as many samples as possible, e-mail providers largely rely on crowd-sourcing. The idea of this crowed-sourcing is that every user of an e-mail provider tells the provider about e-mails that were falsely classified. Therefore, e-mail providers implemented automatic feedback systems for the e-mail filtering mechanism. These feedback systems usually take the form of a button on which a user can click to identify an e-mail as spam, if it reached the in-box, or as legitimate, if the spam filter sorted it in the spam folder [56]. Using those reports, the spam filters can dynamically update their internal filtering patterns, which enables them to react dynamically to new spam campaigns. Without regular updates of those patterns, filtering mechanisms become ineffective extremely fast, because spammers are constantly trying to find new ways for their spam to bypass the filters. An example for such a new trick is purposely misspelling certain words in spam e-mails. This became common practice when spammers noticed, that spam filters were filtering for a certain set of words. Moreover, apart from simply obtaining more samples, the crowd-sourcing approach also generates samples, which describes cases in which the patterns did not produce the correct result. Thus teaching the filter rules for avoiding such false positives or negatives even faster [56].

However, “Spam” and “Not Spam” voting will produce falsely classified e-mails in the training sets. The reason for this is that untrained users are bound to make mistakes in detecting spam e-mails. For example, users regularly report newsletters as spam, even though they subscribed to them. Alternatively, users also make mistakes by mis-clicking the report button. Nevertheless, the amount of mistakes in the training sets is usually so small that they do not interfere with the filters learning process.

But once spammers understood how these reports worked, they found a way to influence the filters by abusing the reporting system. They noticed, that if they created accounts and purposely reported their spam e-mails as legitimate, the time it took the filters to detect their spam mails reliably could be increased. In the study performed by Ramachandran, Dasgupta, Feamster and Weinberger [56] about the abuse of these reporting systems, they found that in a four month long sample taken at one of the major e-mail providers, 51 million “Not Spam” reports were suspicious. A report was considered suspicious, if the reporting user had never reported a spam e-mail. The number of suspicious reports shows that spammers are trying to manipulate filters with fake reports, which makes the amount of valid reports

even more important since only enough valid reports can render the fake reports meaningless.
Even with these efforts of spammers to prevent detection, e-mail providers are today able to generate quite efficient rule sets for labeling e-mails as well as up to date black and white lists of IP and e-mail addresses. Thus it has become harder for spammers to use their own e-mail servers for sending profitable amounts of spam, as those servers would be blacklisted very fast. Additionally, due to the efficient learning algorithms spam filters use, generating fake accounts at large e-mail providers with a home computer has also become less profitable, as those accounts are identified extremely fast and shut down as well. Therefore, spammers had to find ways to become faster than the filters. So they started to use multiple machines at once via the use of botnets.
A botnet is a network of computers on the internet, which perform certain actions on the command of a so called command and control server (C&C). They are usually created by criminals, which compromise multiple machines on the internet, for example by distributing malware or using open vulnerabilities, and install their bot software on the devices. Such compromised systems can be called bots. The bot will then listen to a dedicated server controlled by the criminals waiting for commands. In case of spam senders, the bots will receive a command that tells them what to send and to whom.
This allows spammers to send massive amounts of spam from multiple IP addresses or create multiple accounts with large e-mail providers from which they can send their spam e-mails. It is estimated that today, 2017, 88% of all spam messages are sent through such botnets [76]. Another new strategy they adopted to prevent blocking of their spam e-mails is to hijack legitimate user accounts and use them to send spam e-mails [66]. In a survey performed by Shay, Reeder and Consolvo [66], 30 % of participants, whose e-mail or social media accounts had been compromised, stated that their accounts were used to send spam to their contacts. The reason why hijacked accounts are used is that they will have mainly sent legitimate e-mails up until they were hijacked. Therefore, it will take longer for them to be labeled as spammer accounts, which in term lets them stay alive longer. This allows them to be used to send more spam mails which generates more profit for the spammer.
In order to fight any of these two new techniques the intrusions, which have to have taken place before these accounts could be abused, have to be discovered. To discover such an intrusion either unexpected behavior has to be discovered, or abuse performed by the compromised machine has to be reported. Such reports include for example reporting them as spam sending machines or accounts.
All mentioned mechanisms rely on the speed, at which they detect spamming accounts or servers, to be effective in the long run. Because a faster detection means faster enactment of counter measures, such as closing spammer accounts or black listing spamming servers, which in turn leads to less spam sent per account or server. This results in smaller profit for the spammer. In the best case scenario, the enactment of countermeasures is so fast that the spammer is unable to extract a meaningful profit, putting the spammer out of business in the long run.
The fastest way to detect spam is through automatic filters and end user reports of those e-mails. But since automatic filters can only detect spam messages that share similarities with already reported spam messages, the most important method to fighting spam are end user reports.

### 2.1.3 Phishing E-mails

In Section 2.1.1 some subtypes of spam were mentioned. One of those were phishing e-mails. The *Anti-Phishing Working Group (APWG)* defines phishing e-mails as e-mails which employ both social engineering and technical subterfuge to steal a consumer's personal identity and financial account credentials [37].

Social engineering in this context relates to influencing and deceiving individuals through psychological means. Which psychological methods can be used to influence people and how they are used is discussed in Section 2.1.6.

Technical subterfuge in terms of phishing e-mails refers to malware, which can be either appended to the e-mail directly or distributed as downloads on linked phishing sites in the e-mail.

However, this definition is quite restrictive regarding the goal of phishing e-mails, as only either trying to steal a user's identity or financial information are excepted objectives. But there are several other attack motives which also use social engineering and technical subterfuge to achieve different objectives.

Thus, in this thesis phishing e-mails are defined as e-mails which employ both social engineering and technical subterfuge in order to obtain sensitive information, financial benefits or access to sensitive systems.

Therefore, e-mail campaigns which try to spread malware, such as ransomware or bots, fall under the phishing definition as well.

In order for phishing e-mails to be successful, they have to deceive their victims into believing in the legitimacy of the requests in the e-mail. This is done in two major steps. The first step is to deceive the victim about the source of the message. An example for such a source deception would be a spoofed sender address or a sender address looking very similar to a legitimate address, such as paypa1 appears similar to paypal. This is done in order to impersonate a trusted organization or person and thereby gain the trust of the mail recipient. The source deception is further cemented by choosing the content of the mail to look and relate to the impersonated sender address. A real life example for such a phishing mail can be seen in the phishing mail directed at the Hamburg University of Technology in Figure 2.1. In this mail, the sender appears to be info@tuhh.de, which would appear to be an internal e-mail address. However, when looking at the e-mail source (Figure 2.2), the received header shows that the e-mail did not originate from an internal mail server, but rather from a mail server located at the University of Düsseldorf, the Heinrich Heine University Düsseldorf (HHU).

The second step is to convince the receiver about the validity of the request in the mail. This is performed by presenting the user with made up but plausible emergencies that request an immediate reaction of the user, otherwise some dire consequences will occur. These consequences do usually consist of fines, fees or other forms of financial loss. But as seen in the example phishing mail in Figure 2.1, it can also be something mundane such as loosing an e-mail account. The purpose of the emergency situation lies in creating a sense of urgency [14], which leads the recipient to act upon the mail's request instead of questioning its validity.

The way phishing e-mails are distributed is very similar to spam messages, which means they are sent

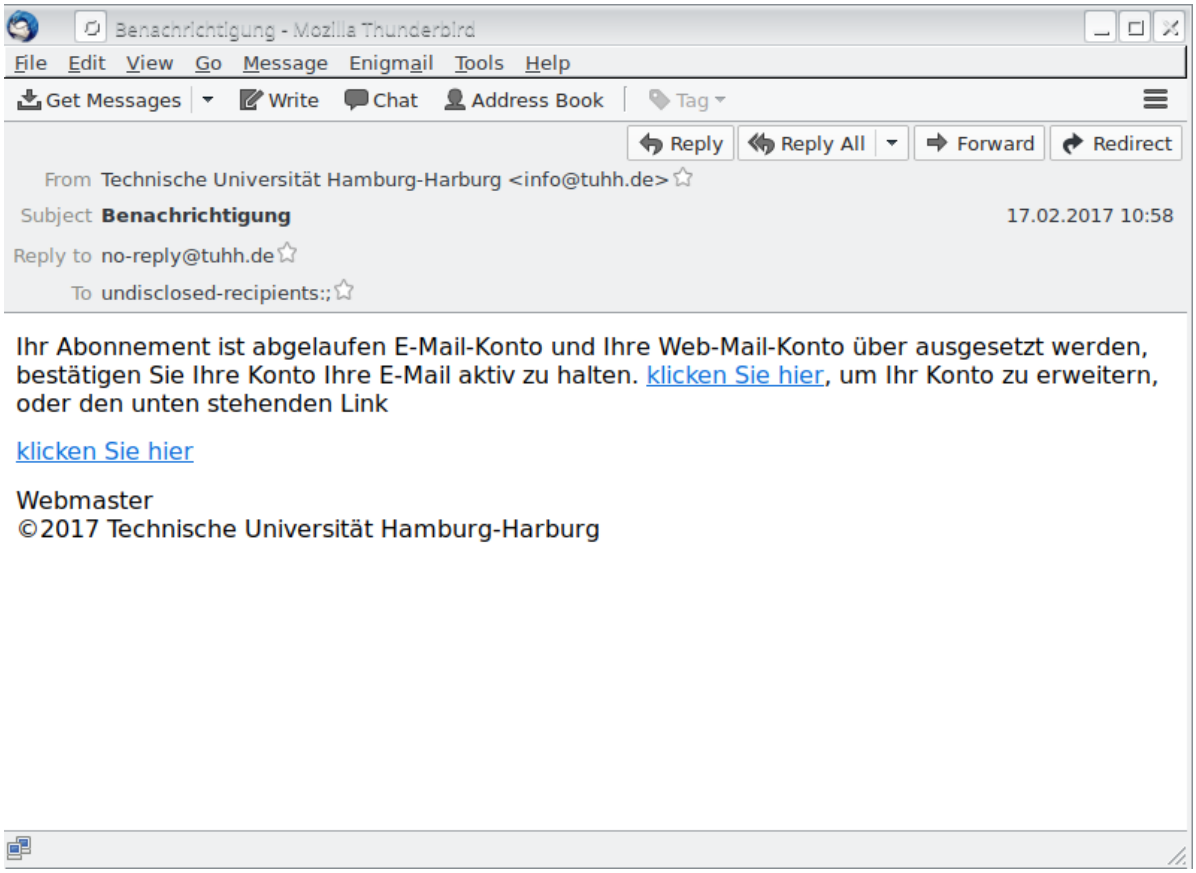


Figure 2.1: Phishing e-mail targeted at a German University

```
Received: from smtp4.rz.tu-harburg.de (smtp4.rz.tu-harburg.de [134.28.205.34])
    (using TLSv1.2 with cipher ECDHE-RSA-AES256-GCM-SHA384 (256/256 bits))
    (Client CN "smtp4.rz.tu-harburg.de", Issuer "TUHH CA in DFN-PKI Global - G01" (verified OK))
    by mail.tu-harburg.de (Postfix) with ESMTPS id 3vPpQP5Q8gzhXyR;
    Fri, 17 Feb 2017 10:59:09 +0100 (CET)
Received: from thor.rz.uni-duesseldorf.de (thor.rz.uni-duesseldorf.de [134.99.128.245])
    (using TLSv1.2 with cipher ECDHE-RSA-AES256-GCM-SHA384 (256/256 bits))
    (Client CN "thor.rz.uni-duesseldorf.de", Issuer "Uni Duesseldorf CA-G01" (verified OK))
    by smtp4.rz.tu-harburg.de (Postfix) with ESMTPS id 3vPpQP0gYBzGmRF;
    Fri, 17 Feb 2017 10:59:09 +0100 (CET)
Received: from roundcube.hhu.de (unknown [192.168.25.1])
    by thor.rz.uni-duesseldorf.de (Postfix) with ESMTPA id D3E30C28A2;
    Fri, 17 Feb 2017 10:59:08 +0100 (CET)
```

Figure 2.2: Received headers from the phishing mail in Figure 2.1

unsolicited to many receivers, using the large number of recipients to compensate low deception rates. But there exists also a subset of phishing e-mails that instead of being sent to masses, are only sent to a small selected group of individuals. These e-mails are called spear phishing e-mails and will be defined and further discussed in Section 2.1.4.
After it has been defined what phishing e-mails are and how they are distributed, it is also worth looking at the methods and goals of phishing e-mails. The main goal behind phishing e-mails is to acquire money for the sender. The first possible way, how this can be achieved is by acquiring and selling the personal information of an individual. Therefore, many phishing mails ask their victims to verify some e-mail, social media or other account by following a link in the e-mail. The link directs their victim to a phishing site created prior to sending the phishing mail, at which the victim is asked to input personal information. This information can among others include username, passwords, addresses, phone numbers but also sensitive information regarding organizations or products, which can then be used by the attacker to launch further attacks, sell them or steal the victims identity for other criminal purposes.
Alternatively, some attackers aim to collect financial details of their victims directly. Those phishing e-mails often impersonate banks or other financial institutions. In these links they ask the recipient to look at some strange occurrences in their accounts by again clicking on an embedded link. This link again links to a phishing site which in this case either harvests the banking credentials directly or installs a trojan that spies on the banking credentials of the victim [77].
Another way the attacker can get his profits is by extortion. Those e-mails usually claim to contain a bill that is overdue or similar documents as attachments. However, the appended file does not contain a bill, but rather malicious software. An example for such an attack is a ransomware attack, in which ransomware software infects the victims machine and encrypts data. In order to regain control over his or her machine and retrieve the data encrypted by the attacker on the machine, the victim has to pay the attacker for a decryption key [78].
Last but not least some phishing mails use attachment or browser vulnerabilities to install a backdoor on the victims system system. The backdoor can then be used to integrate the victims machine into a botnet.

### 2.1.4 Spear Phishing E-mails

Most phishing e-mails are not focused on specific individuals but rather towards the mass audience. Therefore, these e-mails cannot be designed for each specific recipient, but rather to appeal to anybody. This obviously leads to a lower success rate, because people expect organizations they trust, such as banks, retailers and so on, to know certain personal detail. However, this disadvantage is compensated by sending the same phishing mail to a huge amount of recipients. Thus even if the success rate would lie below 5%, the sheer numbers would result in profits for the attacker. However, sending phishing e-mails to massive amounts of recipients does not work if the target of the campaign is to obtain specific information known only by a small group or even one person [36]. Therefore, phishers started to design more targeted phishing mails. These targeted phishing e-mails are called spear phishing e-mails.
In order to craft their phishing mails specifically for their target, they have to gather information about

their intended victims. Thus they perform research about them, by searching through publicly available curriculum vitae or by investigating social media accounts. The information that is gained is then used to create personalized believable phishing e-mails for their victims. These e-mails contain content that was selected to relate to the potential victims and contains personal information. This increases the probability of successfully deceiving the victims [54]. At the same time this increases the effort that has to be put into the phishing e-mail. Therefore spear phishing e-mails require five times the investment of regular phishing e-mails. However, due to the precise target selection and increased success ratio, it is estimated that spear phishing e-mails can yield 10 times the profit.

To compare the success of regular phishing and spear phishing, an experiment at the Indiana University was conducted [41]. In this experiment university students between the age of 18 and 24 were divided into two groups. The first group consisted of 94 students and will be referred to as control group. This group received a phishing e-mail from the researcher, which appeared to be coming from an unknown (non existent) student and contained a link which appeared to link to an university internal location. In this group about 16% fell for the phishing mail. The second group consisted of 487 students. They received the same e-mail, however this time the e-mail appeared to originate from one of their friends or acquaintances. The information about friendships and acquaintances was collected beforehand by data mining relationship information from multiple social networks. All the information was freely available to everyone and the researchers used only publicly available tools to aggregate the information. This new targeted spear phishing mail was far more successful than the original phishing mail. From the 487 students a total of 349 fell for this mail, which equates to 72%.

Furthermore, looking at the reactions to this experiment, the researchers noted that victims of the spear phishing campaign felt that their privacy was harmed. This stemmed mostly from the mining of relationship information. The students complaining were unaware that their private information, which they posted into social media accounts, was accessible by everyone. Therefore, many believed to have been hacked. This underlines, that many users vastly overestimate the effort attackers have to perform in order to obtain personal information, such as relationships to other individuals, which again increases the susceptibility to spear phishing attacks.

### 2.1.5 Efficiency of phishing e-mails and websites

Up until now phishing and spear phishing have just been described, but it has not been stated how dangerous they are. To judge the risk phishing e-mails pose for individuals and organizations, several studies and examples are analyzed.

The first study that is looked at is the *Data breach report 2016* published by Verizon [68]. The Verizon data breach report 2016 is an analysis of 64.199 incidents and 2.260 data breaches that were either found by Verizon during paid forensic investigations or reported to Verizon by contributors. These incident and data breach samples contained 9,576 total incidents from which 916 were confirmed as breaches, meaning they disclosed data to unauthorized parties. This showed that about 12% of all phishing attacks were at least partially successful.

In another study [50], the susceptibility to phishing in an university setting was researched. To evaluate

this, the researchers sent phishing e-mails to the unsuspecting participants of the research, trying to lure them into revealing sensitive information. In the first test run the researchers sent non personalized e-mails, meaning each participant received the same e-mail, word for word. Looking at the results of this first phishing campaign revealed that nearly nine percent of the participants opened the e-mail, clicked on the link and provided valid passwords at the phishing page.
The 3% difference between the success rates can be explained by the difference in the used data sets. The study about phishing at a university did only look at regular, non personalized, phishing e-mails, while the data breach report includes also spear phishing e-mails in their data set about phishing incidents. Furthermore, the university study did solely include phishing e-mails which directed their victims towards phishing sites, the data breach report did also contain phishing e-mails that spread malware.
Malware distributing phishing mails most of the time require only one successful deception, which leads to opening an attachment or visiting a site, which automatically uses vulnerabilities to install malware. In contrast, phishing e-mails trying to obtain information most often require two successful deceptions to achieve their goals. The first deception is necessary to lure the victim into visiting the phishing site through a link. However, this phishing site now has to deceive the victim again in order to make the victim believe to be on the legitimate site, such that the victim feels comfortable to give away the sensitive information. Therefore, these phishing websites give potential victims another chance to detect the deception.
Thus, in order to evaluate the threat of phishing e-mails in a meaningful way, the successful deception rates of phishing websites have to be evaluated. In the study, *Why Phishing Works* [21], the researchers asked 22 participants to classify 20 websites. The websites consisted of seven legitimate sites, nine representative examples of real world phishing sites and four sites crafted by the researchers to evaluate certain phishing techniques. The participants were able to interact with these sites freely and use any method they deemed useful to help them identify phishing sites. The researchers found, that the deception rates of the phishing sites ranged from 18 to 91%. Thus they concluded that a well made phishing site could deceive over 90% of visitors. However, the researchers also tried to identify why people fall for phishing attacks. Therefore, they looked through the data inside the *Phishing Archive* [35] maintained by the *Anti-Phishing Working Group* and were able to identify three reasons why people are deceived by phishing attacks [21].

**Lack of Knowledge** The first reason why people fall for phishing websites or e-mails, is a lack of knowledge about computer systems and security indicators. The lack of knowledge about computer systems is widely exploited by attackers. For example most computer users do not know how e-mail communication works in detail. Thus, they do not know about e-mail headers and how they can be spoofed. Therefore, an attacker spoofing the *From* header, such that the e-mail appears to originate from a trusted source, is likely to deceive users into trusting the e-mail. Furthermore, phishing sites often use the lack of knowledge about URLs to deceive people. For knowledgeable individuals it is quite easy to identify the domain and sub-domains of an URL, but most end users cannot. Therefore, many believe that URLs such as *www.paypal.com.evil.net*

lead to paypal.com and not as in reality to evil.net.

The other field in which end users lack knowledge are security indicators. For example most modern browsers contain a visual representation whether or not a connection is encrypted. However, most end users never bother to hover over the small lock icon which indicates the encrypted connection and thereby never get to know what this icon means. But even if they know what an indicator means, many users do not know where and how they should be displayed. Thus they are easily fooled by supposed security indicators appearing in the web page instead of the browser.

**Visual Deception** The visual deception characteristic describes the deceptive capabilities of phishing websites and e-mails. These largely rely on the appearance of the site and/or e-mail. However, there are three different types of visual deceptions.

The first type of visual deception are deceptive texts. Deceptive texts employ different strategies to deceive the user into reading the text differently than it is written. For example some letters in links can be replaced in such a way that unobservant people will read the text wrong. Such a replacement would be to switch the letter l with a 1 (one).

The second form of deception is the deception through images. Images can replace text in e-mails and phishing sites and thereby mask links or catch clicks that were supposed to select texts.

The last category is the deceptive look and feel category. This category describes the visual deception phishing e-mails and websites achieve by being designed similar to, or even copying the legitimate sites design.

**Bounded Attention** Even if individuals have the required knowledge about security features as well as computer systems, they still fall for phishing mails and sites. The reason for this is a general lack of awareness. Many users neither tend to notice the security indicators nor the absence of any of them, nor read warnings.

The effects of those weaknesses of individuals can be further increased if the phishing mails exploit other psychological vulnerabilities through social engineering.

In another paper *Why phishing still works* [4] an experiment was performed in which 21 participants were again asked to classify 24 websites, as either legitimate or phishing. Among these 24 websites were 10 legitimate websites and 14 phishing websites. The participants were presented with the website and could freely interact with the page, just as if they would have opened it in their browser. The phishing sites could be detected through missing security indicators, false URLs as well as other mistakes in the sites design. The results of the experiment showed that the participants were able to identify between 9 and 22 sites correctly. Thus each participant was unable to classify between 8% and 62% correctly. The average classification error rate was 37%. It has to be noted that this number includes false positives and false negatives. Looking at just the phishing sites that managed to be classified as legitimate sites, results in a deception rate of 47%.

The success rates are however idealistic approximations, since the participants were explicitly tasked with classifying the websites. Hence, they were much more aware as in their daily routine and checked

the security indicators known to them more carefully.
After analyzing how successful phishing e-mails are at deceiving people and why they are as good as they are, it is important to assess the damage potential of phishing. Since the goals of phishing attacks are so diverse and the phishing has the ability to overcome any security measure, the damage potential is nearly infinite. Those damages can come in any form, from monetary and data losses to the loss of reputation. The best way to get an overview of potential costs for victims is to look at some real live examples.
The first example for a successful phishing campaign can be seen in the large successful ransomware attack wave against health care organizations [47]. In this campaign the organizations received phishing e-mails with either a malicious word document or javascript payloads, which loaded and installed the locky ransomware. Locky encrypts the files on the infected system and demands payment for the decryption key. In this campaign Intel Security estimated, that at least 100.000$ in ransom were paid.
Another prominent example is the incident at Epsilon in 2010 [38, 8]. In this incident customer data of multiple Epsilon clients was accessible by the hackers. Epsilon is the worlds largest e-mail provider with 2500 clients globally, among them are companies such as the citigroup and Hilton Hotels. The hackers were able to compromise Epsilons security by sending specially crafted phishing e-mails to employees with the desired access rights. In those e-mails were links to phishing sites, which downloaded and installed multiple types of malware on the machines and gave the hackers access to the mailing list information. The incident resulted in a loss of reputation as well as subsequent costs for Epsilon and their clients.

### 2.1.6 Social Engineering and Phishing

There exist many different definitions of social engineering, which mostly disagree on whether this term can be applied to technology based communication or not. In this thesis, the definition by Evans [25] will be used. He defined social engineering as the exploitation of human vulnerabilities in organizations or systems, by psychological means. Social engineering according to his definition has to have the goal of either obtaining knowledge or access to a system.
According to this definition, every phishing e-mail falls under social engineering, since each and every phishing e-mail attacks the human element in security. However, the definition does not explain where the vulnerabilities lie and how they are exploited.
In the fields of psychology as well as marketing a lot of reasearch has been done regarding factors which influence our behavior. Therefore, many findings of marketing research are applicable to social engineering as well. During the analysis of different marketing and advertising strategies, Cialdini [15] was able to extract six basic principles of influence.
Concurrently, Gragg discovered seven psychological triggers [34], which help social engineers deceive their targets. In 2008 Scheeres discovered that the seven triggers and Caldini's six principles were describing similar behaviors, but from different points of view and thus was able to link Caldini's six principles to the psychological triggers [63]. Thereby, he implicitly showed that Caldini's principles can also be used to describe the human vulnerabilities in social engineering and thereby phishing.

This was further researched by Ferreira, Coventry and Lenzini in their study [26]. By looking at the different persuasion principles and phishing e-mails, they were able to show that the persuasion principles were largely used in phishing e-mails. Moreover, they were able to show that most phishing mails employed multiple principles at once.
The six principles are each identified by a keyword and describe a specific principle that influences people. These principles are authority, commitment and consistency, liking, reciprocity, scarcity and social proof. In the following a closer look is taken at the individual principles.

**Authority:** In situations in which people are insecure or uncertain how to act, they tend to look at the behavior and opinion of higher authorities and follow their example or advice. This behavior was identified by Caldini as the authority principle. He defined it as the principle that people are very likely trusting statements and recommended behaviors issued by supposed experts. An example for this can be found in advertisements with scientific recommendations such as *"9 out of 10 dentists recommend"* in toothpaste [17] and toothbrush commercials. Those commercials impact our buying decisions, but they are not utilizing the full effect of the authority principle. The reason for this is that the experts are perceived as biased, because they are obviously paid for the adverts. However, it is easy to imagine that if the chosen dentist of an individual, with no clear connection to the company, is issuing the recommendation in private to his or her patient, the impact would be a lot higher.
Furthermore, individuals tend not to question supposed experts, even if their own judgment contradicts the experts opinion, as long as the expert appears to possess higher authority and is either independent or his or her interests are clearly aligned with the individuals interests.
A particularly troublesome example of this is the so called *Captainitis* [30] in aviation. This phenomenon appears in situations in which the co-pilot notices some strange and for him potentially serious incident or misjudgment and notifies the captain about it. However, if the captain then dismisses the argument, the co-pilot will not follow up on it and will not disagree with the captain. This can lead to potentially catastrophic consequences such as plane crashes.
This phenomenon occurs due to the huge authority boost the position of captain is providing, which makes it easier for the captain to dismiss information or opposing views. At the same time the authority difference between the co-pilot and the captain hinders the co-pilot in defending his point of view as it would mean to directly infringe the captains authority.
However, this principle does not only find application in advertising and group dynamics, but additionally in social engineering and phishing. The simplest application in a social engineering attack is for the social engineer to pretend to be of a higher authority, such as a member of management or team leader, inside of the targeted organization than the victim. Thereby advising him or her to perform the actions he or she wants. Most organization members will not query the validity of the tasks, as long as the social engineer manages to remain a convincing authority figure.
Since in phishing e-mails there is no face to face contact, the attacker has to rely on authority


*February 22, 2017*

**Secure Documents**

Please find attached your secure documents. Please review, complete and return completed documents via email to CRAoffice@cra-arc.gc.ca.

If you have any queries relating to the above, feel free to contact us at: CRAoffice@cra-arc.gc.ca

Confidentiality Note: The information contained in and transmitted with this communication is strictly confidential, is intended only for the use of the intended recipient, and is the property of Australian Taxation Office or its affiliates and subsidiaries. If you are not the intended recipient, you are hereby notified that any use of the information contained in or transmitted with the communication or dissemination, distribution, or copying of this communication is strictly prohibited by law. If you have received this communication in error, please immediately return this communication to the sender and delete the original message and any copy of it in your possession.

Figure 2.3: Phishing E-Mail posing as an E-Mail from the Taxation Authority [10]

derived from titles or position. For example an attacker can impersonate governmental agencies, such as the taxation administration. This can be seen in Figure 2.3. In this example e-mail, the appended word document contained malicious macros which would download and install malware on the victims machine. Even though the phish is not that well crafted, note the confidentiality notice at the bottom mentioning that the information in the e-mail belongs to the *Australian Taxation Office* even though the mail comes from the Canadian Revenue Agency, it is still very potent and uses the authority of a governmental institution to legitimize the request for an action.

Alternatively, it is possible to impersonate organization internal authorities such as the administration or supervisors. In the analysis performed by Ferreira, Coventry and Lenzini [26], they found that among phishing attacks the authority principle was the third most used principle, in incidents related to data theft and malware.

**Consistency and Commitment:** When people make decisions, they tend to try to remain consistent with prior decisions and actions. Therfore, according to this principle, if someone agrees to sign a partition or display a sticker promoting some cause, he or she will be more likely to agree to larger requests regarding the same topic. This hypothesis was scientifically reseached by Freedman and Fraser in their study *Compliance Without Pressure* [31], who found strong evidence for the validity of the initial hypothesis. Furthermore, they found that this effect is noticeable even if the two requests are not similar at all or relate to different topics.

This principle is widely used in our everyday lives, for example in fund raising for humanitarian causes. To maximize the funds they are able to collect, fund raisers will at first ask someone if he has two minutes to spare for the cause. This marks the first request. Then the fund raiser is giving a short speech about the topic. After this, the individual listening has already committed to invest time to listen to the cause. Thus they are more likely to donate when the fund raiser asks for a donation for said cause. Another example is distributing fliers to people. In this case the individuals are asked to accept a flier for an organization, for example a food delivery service. Even taking this flier is a small initial commitment, which makes the individual more likely to test the advertised services. However, since in both cases the initial investment is very small and no strong stand has been made, the effect of the consistency and commitment principle is quite small.

Nevertheless, social engineering schemes often employ this principle. As an example many schemes, which rely on personal face to face communication, start by asking for some unimportant and most often not very confidential information or a small favor. After the first contact, the requests become larger and many people will not hesitate to fulfill them as well. However, without the first small requests and commitments, they would not have fulfilled the larger requests or at least questioned them.

In phishing mails however this principle is rarely used. Those phishing e-mails that are sent to a huge number of potential victims, which the attacker does not know, do not want to make multiple small requests in succession before getting to their goal. Such a sequence of requests would provide a potential victim with more time to think about the conversation and thereby increase the detection probability. Furthermore, the phisher would need to utilize the same impersonated personality with the same valid e-mail address to contact the victims or make all the requests in one mail. Thus either e-mail account suspensions would become more harmful for the attacker, or the single e-mail would have to be longer and thus it would be easier to see through the deception.

Spear phishing e-mails are different. Before sending spear phishing mails an attacker usually researches about the victims. Therefore, he or she can craft the phishing e-mails specifically for the victims, lowering the detection possibility significantly. Furthermore, since the amount of potential victims is rather small and the goal of such an attack is strictly defined, detection or the phishing scheme usually equates to an immediate failure anyway. Thus account suspensions due to detection do hardly matter. Additionally, the increased potential to successfully deceive an

individual, that comes with the consitency and commitment principle, make an extended communication more viable for spear phishing attempts.

**Liking:** The liking principle is based on the natural desire of humans to trust and comply with requests made by people we personally like. The reason why we like a specific person are unimportant for the efficiency of the principle. For example physical attractive people and people with similar interests are both equally efficient in influencing us, because psychologically humans tend to not differentiate between the reasons why we like someone.

An example in the field of marketing can be found in the tipping behavior towards waiters. For them it is quite important to get the most tips from their customers in order to earn their living. Thus, the researchers Rind and Bordia studied how the drawing of a smiley onto the receipt would influence the tips given by customers [60]. In the study they found that waitresses, who draw the smiley, were perceived as friendlier and more sympathetic. Thus customers increased their tips by roughly 19 percent. On the other hand, male waiters drawing the smiley were not perceived as friendlier, but this behavior was rather seen as strange. The reason for this is that society sees such expressive behavior as untypical for males. This perception of strange behavior of the male waiters actually led to a decrease of tips for them.

The study thereby showed that the liking principle is quite effective, as a 19 percent tip increase for the waitresses is a substantial boost. However, the likability increase or decrease of certain action is largely effected personal aspects, such as gender and culture.

Obviously this principle is abused by social engineers, since appearing likable is not a difficult task to achieve while giving a large reward. For example simply dressing appropriately and acting nice, can lead to trust and support of unsuspecting victims.

In spear phishing e-mails this can be used as well. If the attacker plays the role of a new employee who has a problem and requires assistance, many friendly colleagues will try to help. During this assistance, they are, due to the previously mentioned consistency and commitment principle, more likely to give away confidential information than they would normally do.

In other attacks, in which masses of phishing e-mails are sent however, the liking principle is less useful. The reason for this is that an attacker usually does not invest time into building a sympathetic image of him- or herself, but rather relies on other influencing principles.

**Reciprocity:** In human societies favors play an important role in our everyday lives. Generally we as humans tend to try to repay favors, thereby opening our selves up for manipulation.

Examples of quite harmless manipulation attempts can be observed in restaurants every day. A simple tacktic is to give the guests a small gift, for example a small piece of chocolate, at the end of their meals. This has proven to increase the amount of tips which are paid as well as the overall satisfaction of customers. In a study Regan [59] showed that a salesman for raffle tickets could increase the number of tickets he can sell, if he hands the potential customers a coca-cola can for free beforehand. The gift the recipients received increased the sales regardless whether the gift was related to the sale or not. Moreover, the increase in sales even generated a higher


Figure 2.4: Detran Malspam E-Mail [22]

profit than the initial investment, thereby allowing the salesman to increase his profit.

This principle has also a very common usage in social engineering attacks. For example a social engineer who calls an employee of a company could claim that there is a problem with the internet and that he needs to disconnect said employee for a certain, quite long, amount of time. This will most likely lead to the employee protesting, because he will be unable to work during this time. Then the social engineer can offer the employee the supposed favor, that he or she will be notified when exactly his connection will be terminated and that the time span he or she is disconnected will be kept as short as possible. However, to do so, the social engineer needs the employee's user name for this. Later the social engineer then calls the employee again, in order to tell him or her that he or she will be disconnected and ask him or her to log off. Additionally, the social engineer asks the employee for his or her password, which he or she claims to need to test the changes and thereby get the employee back online faster. Many employees will comply with the request for the password as well as the user account. Especially since they are currently receiving a perceived favor, thus they do not question the requests. Such an attack was described by Kevin Mitnick in his book, *The Art of Deception* [49].

Apart from social engineering via the telephone, this principle can also be used in a slightly modified version by attackers in their phishing e-mails. For example in 2016 a phishing campaign in Brazil claimed to come from Detran, which is an abbreviation for *Departamento Estadual de Trânsito*, the Brazilian institution for ground vehicle supervision. The e-mail, see Figure 2.4, contained a supposed ticket for crossing a red light. However, in the message was a passage, in Portuguese, which claimed [22]:

*You can download the application of the Electronic Notification System (SNE DETRAN), which allows you to receive notifications at any time and place, also offering a 40% discount on the fine for payment until the expiration date of the same.*

The links in the e-mail then led to a download site for malware. The 40% discount on the fine counts in this specific incident as the favor, because it is a considerable discount provided to the victim basically for free. Thus, even if the recipient is skeptic towards the traffic violation claim, made by the phishing mail, he or she might follow the link in order to either clear up the confusion or at least get the promised 40% discount.

Another form of the reciprocity principle is reciprocation for declining a request. For example, if one person makes a huge request to another person, which declines the request, this person will feel the need to reciprocate to the requester. Therefore, he or she will agree more easily to the next smaller request made by the requester. This was examined by Cialdini et al. in the study *Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique.* [16]. In this study the participants were split into two groups. The students in the first group were asked whether they would chaperon juvenile detention center inmates on a field trip for one day. From these students only about 17% agreed. The other half was then approached and asked whether they would tutor said inmates for two days each week over a period of two months. As expected, the agreement rate for this request was much lower. However, after denying the first request the students in the second group were asked whether they would then chaperon the inmates on the one day field trip. From those students, more than twice as many students agreed to this request as did in the first group.

In face to face social engineering attacks, this principle can be used to obtain information. An example would be a social engineer who asks for an information, that is not available to the victim. The victim then has to decline the request, which the social engineer uses to ask the victim for a related confidential detail, claiming that this would enabling them to at least partly fulfill their task.

In phishing e-mails however, reciprocation for denying requests is mostly unused, as it would require more than one contact. Therefore, it is only viable for spear phishing attacks. However, denying the first request could lead to doubts regarding the legitimacy of the request, thus making using this part of the principle quite risky.

**Social Validation:** Humans are highly socialized. Thus people who are unsure how to act or react try to align their behavior with the majority. This psychological principle is responsible for the creation of social norms, which Caldini called the social validation principle.

In advertising, this principle is widely used, for example when Mc Donald's claims "Billions and Billions served" in their slogan. This slogan suggests that it is normal for people to eat at Mc Donalds. Thus if you do not know or are unsure what to eat, go to Mc Donalds. But Mc Donalds is not alone, many insurance companies use the number of customers they have in their advertising. This is supposed to influence customers looking for a new insurance company by

asserting quality through their number of customers. Moreover, the social validation principle has proven to be effective in a variety of other applications, where human behavior could be influenced. One such occations was a field study on curbside recycling [64]. In this study, normative feedback was used, to educate people about recycling, with the long term goal to make them recycle more. The idea behind the normative feedback is that feedback, which creates and emphasizes the social acceptance, will make recycling a social norm. Thereby the target group should be influenced to change their behavior and recycle more. In their study normative feedback, explicitly establishing recycling as a social norm, increased the amount of participation in the recycling process.

Using the social validation principle in phishing requires a little bit more effort, as complying with the request has to be either by default socially accepted, or the social engineer has to create an environment in which complying with the social engineers request is a social norm. One of these environments for giving information to strangers is the new employee situation. Thus a social engineer who impersonates a new employee can extract a lot of information about internal processes, company language and internal structure without creating suspicion on the other side, simply because it is in many companies a social norm to answer questions regarding such areas from new employees.

Phishers who utilize the social validation principle can use it in a similar way in their phishing mails, by simply asking questions in spear phishing mails directed at an organization. However, a detection is more likely for them as they either have to use an organization e-mail address, fake the origin of the e-mail, or use an e-mail address so similar to the organization addresses that the unobservant user does not notice it. This is therefore usually only done in spear phishing campaigns.

However for large, non-targeted phishing schemes, the social validation principle can still be used. For example phishers can try to impersonate a charity organization and ask for donations to help in a current crisis, or to participate in a e-mail partition for some cause. In both cases links in the e-mail lead to phishing sites on which malware can be downloaded, or user information can be stolen.

**Scarcity:** Caldini's scarcity principle is based around the human desire to acquire scarce resources. Additionally, humans believe scarce resources to be more valuable than other more common resources.

The prime example for this is the desire for diamonds. People are willing to pay huge sums of money for diamonds, while their main use for most people is purely decorative. Thus the price of diamonds simply depends on the scarcity of the resource. This desire for scarce resources is widely utilized in discount campaigns and the introduction of new products. For example when Apple releases a new iPhone, the quantity that is available at launch is much smaller than the predicted demand. Thus people are camping outside of Apple stores, generating a hype just because of the scarcity of the new product, that otherwise would not develop.

Social engineers can use this principle as well, by claiming to perform a short questionnaire that gives a reward to the first 100 participants. Thus people are compelled to not think about the offer but rather complete the questionnaire as fast as possible, giving away confidential information in the process.

In phishing e-mails this principle is used even more. The reason for this is that scarcity always creates a feeling of urgency, thus helping with the deception of potential victims. An example for this are phishing e-mails which claim to contain an overdue notice for some bill, which will lead to legal actions if not paid immediately. The scarcity of time to prevent legal measures in this instance will make recipients look at the attachment. Thus, upon opening the attachment, malware can be installed and the attacker reached his or her goal.

Even though the principle above are described separately, in the real world of phishing these principles are used in conjunction with each other. Such that a phishing e-mail uses the authority principle, by impersonating a big company or an debt collection agency, and the scarcity principle, by claiming the victim having only a few days to act before judicial measures will be taken. This combination makes attacks even more successful.

### 2.1.7 Individual Susceptibility to Social Engineering

But not every user is equally effected by the different principles. Their susceptibility to phishing depends on their personality traits as well. In their liturature study, *The Social Engineering Personality Framework* [80], Uebelacker and Quiel stated correlations between the susceptibility to the different principles of influence and the Five-Factor Model (FFM) for personalities. Thereby correlating the susceptibility to the individual influence principles and the personality of an individual.

The five factors in the FFM model are Conscientiousness, Extraversion, Openness, Neuroticism and Agreeableness.

In their paper, they state that a person with a high degree of agreeableness will be more susceptible to the principles of authority, reciprocity, liking and social proof, thereby making this person ultimately more susceptible to all social engineering and phishing attacks in general. This stems from the fact that people with a high degree of agreeableness tend to be easier to convince and thus easier to deceive. This becomes even worse, if the person making the claim appears to possess a higher authority than the agreeable person.

On the other hand people who lean to the neuroticism trait, are a lot more cautious with their actions. Therefore, they are suspicious of most claims and requested actions. Making it a lot harder to deceive such an individual, which leads to a lower susceptibility to all sorts of social engineering attacks.

Conscientious individuals are more likely to follow rules and regulations. Thus, it seems that this trait makes someone more resilient against social engineering as well. But on the other hand, a conscientious person is also more likely to honor established hierarchical structures, which makes him or her vulnerable regarding social engineering attacks using the authority principle. Furthermore, since conscientious individuals are also more likely to follow social norms, they are also more likely to be influenced by the reciprocity and the consistency and commitment principles. Both of these principles

exploit the social norm of either returning favors or remaining consistent with previous views and actions, which conscientious people care about.
The openness trait in this context means to be open to new experiences. This appears to be quite an undesired trait regarding resilience to social engineering, because it comes with a more vivid fantasy and thus makes individual more likely to be deceived. However, openness to new experiences often implies a higher degree of technological expertise, as individuals with this personality trait are more likely to be curious about how things work. But they also react strongly towards things that appear to restrict their freedom, such as time limited offers or access restrictions to a small amount of people. Therefore, the openness personality trait makes their owner more susceptible towards social engineering attacks using the scarcity principle.
Last but not least, the extraversion personality trait among others describes the sociability of an individual. This sociability also decides how much impact liking and the behavior of others have on our behavior. This influence becomes larger and larger the more sociable we are. Thus increasing also the susceptibility towards social engineering attacks, which utilize the social proof and liking principles. At the same time, extraversion decreases the importance of consistency and commitment for an individual, thereby lowering the susceptibility regarding this principle.

### 2.1.8 Fighting phishing

The previously described studies and examples have shown that protection against phishing mails is important. Such a defense always consists of two elements, educating the individuals and automatic detection.
Educating e-mail users how to detect phishing e-mails is straight forward targeting the problem. However there are a few challenges. The first problem is that efficient training can only be provided in organizations. The reason for this is that most users do not want to educate themselves regarding security behavior [44]. Thus, most organizations have to educate their members themselves. Furthermore, the training has to be renewed regularly as otherwise advice is forgotten and changed attack patterns can not be studied.
Furthermore, such awareness and detection training does not remove the risk of successful phishing attacks, it can just reduce it.
There exist two basic training methods. The first one is lecture based, in which the trainer shows the trainees what to look out for and explains the theory around e-mail and web security. The second method is called embedded training [13, 73, 44]. Instead of a limited amount of training sessions, embedded training is performed over a certain period during the daily routine. The proceedings during such a training are as follows. The trainees receive among their regular e-mail traffic additional phishing e-mails selected by the trainers. These phishing e-mails are not marked in any way and therefore look just like regular phishing e-mails would. However, the phishing mails used for the training do not contain malware or direct the user to a malicious site, rather the links inside of the mail direct potential victims to an information site about phishing and the specific features of the mail that can be used to identify the mail as a phishing mail. This is supposed to raise the trainees awareness regarding phishing

as well as teach him or her how to spot phishing e-mails. But while studying the effectiveness of such a training Escher and Köpsell [73] found, that only about 50% of the trainees read the information page and thus could be educated. However the researcher both assume that even though only half of the trainees who fell for a phishing e-mail got educated, the raised awareness regarding phishing is enough to improve the security against phishing attacks significantly.
In another study regarding the efficiency of embedded training in an organization, Caputo, Pfleeger, Freeman and Johnson [13] sent three rounds of training phishing mails to their trainees and observed their improvements in detecting these phishing mails, depending on the design of the training page they received. Thus, they split their trainees into five groups. Every group received the same phishing e-mails, however the training page these e-mails lead to were different. The first group would just be notified that they had followed a phishing link. In contrast, the other groups would receive a training page which informed them about features in the phishing mail that could have made them suspicious, as well as an explanation why looking for these features is important. Depending on the group, these explanations would be formulated to focus on one of the four perspectives:

1. By following the advice, you can protect yourself from harm.
2. By not following the advice, you put yourself at risk.
3. By following the advice, you protect your co-workers from harm.
4. By not following the advice you put your co-workers at risk.

Looking at the results of their study, they found that a group of trainees did always click on the phishing links, regardless of the type of training they received. They called this group, the group of all-clickers and identified 11% of their trainees as belonging to this group. For this group it did not matter what kind of training they received, they would always follow the links in their phishing e-mails. On the other side, many employees (22%) never followed any of the phishing links at all.
Among those, that were affected by the training, the phrasing regarding the training page proved to be insignificant. However, during their three phishing phases they found an interesting behavior. Among those that clicked on the link, a substantial group noticed directly that they had clicked a phishing link and upon receiving on the training page immediately left the page without reading it. The reason for this behavior was that they thought that the training page was another attempt to fool. Others closed the training page immediately out of shame for falling for a phishing mail. Thus again not all participants were able to receive the desired training.
In an interview after the three phases, the researchers got feedback regarding the training pages and phishing mails. This feedback showed, that generating a training page that does not look threatening and at the same time compresses the meaningful information into texts short enough to be read by the trainees is not trivial. Furthermore, even though the majority of participants believed the training to be more effective than the annual theory training they received prior, some participants in the group that clicked on all links displayed anger towards the training for deceiving them.
These two examples of anti-phishing training show that it is extremely difficult to construct an effective

training regarding phishing detection. Furthermore, due to this difficulty as well as the observed all-clicker behavior, it appears to be impossible to achieve a 100% reliable phishing detection rate through training. Therefore, any defense against phishing has to try and limit the potentially successful phishing e-mails that reach individual humans as much as possible, by utilizing the second defensive tactic, automatic phishing detection.

First and foremost, phishing e-mails are filtered through the same filters as other spam messages are. Therefore, phishing e-mails sharing similarities to other forms of spam are mostly taken care of. However, some phishing e-mails, especially spear phishing mails, evade the regular spam filters very efficiently. The reason for this is that by the nature of their design, phishing e-mails usually try to mimic legitimate communication as closely as possible. For example a phishing campaign trying to obtain financial credentials will most likely send e-mails which are very similar to legitimate e-mails from banks. Thus spam filters are bound to miss at least a few of those phishing mails. Nevertheless, there are some characteristics, that automatic filters as well as humans can search for to detect phishing e-mails [52]:

**Hyperlinks displaying a different target than they are linking to:** In phishing e-mails hyperlinks often display destinations that do not match the destination, that is linked to. For example the following HTML link *<ahref="www.evil.com"> www.bank.com/login</a>* appears to be linking to www.bank.com, but instead it links to www.evil.com. This is often used to deceive people about the location that is linked to. In legitimate e-mails such deceptive links occur only rarely. Thus making such deceptive hyperlinks an indication for a phishing e-mail.

**Hyperlinks linking to IP addresses:** In legitimate e-mails links will nearly always link to registered domains. Thus links that link directly to IP addresses are very suspicious, especially if they display text to the user instead of the IP address they are linking to.

**Hyperlinks displaying text that has no connection to the linked location:** Furthermore, links which display text that is not connected to the location the link points to are also extremely suspicious, since those are also regularly used to deceive recipients. An example would be the link *<ahref="www.evil.com"> click here to claim the money</a>*, which simply displays a message, not indicating at all where it will link to. Therefore, such links do appear quite often in phishing e-mails, making them a potential phishing link indicator. However, many legitimate e-mails do this as well.

**Hyperlinks hidden behind images:** Moreover, if hyperlinks are displayed as an images instead of texts, they can also hide the location a link links to, as well as serve as click-bait to unsuspecting users. Which led to their widespread use in phishing e-mails. Thus the existence of such links is a small indication of an e-mail being a phishing mail.

**Hyperlinks with multiple sub-domains:** Another characteristic in phishing mails is the existence of links with multiple sub-domains such as *www.bank.evil.com*. These sub-domains can deceive people into thinking the site to which the link is directed, is the first sub-domain instead of the

last sub-domain. Thus, these links are often used in phishing mail and can serve as an indicator for phishing.

**Hyperlinks which display multiple domains:** A more extreme form of the previous indicator are links which appear to contain multiple domains. Examples for such deceptive links are *paypal.com.evil.net* and *evil.net/paypal.com*. Both can fool an individual to believe that the links link to paypal.com when in reality these links lead to evil.net. Due to this deceptive capability they are used in phishing e-mails while being very rare in legitimate e-mails.

**Hyperlinks with very long URLs:** Additionally phishing e-mails also contain links with very long URLs. This is supposed to confuse the victim and obfuscate the location the link is pointing to. Legitimate e-mails however, usually try to keep the links short, such that readers will no get confused.

**Pictures loaded from domains different to the hyperlinks in the e-mail:** Most legitimate e-mails contain only pictures that are loaded from the same domain which the hyperlink points to. This is usually the company or event website. However, some phishing e-mails use images from legitimate websites in order to make their e-mails more believable. Therefore, they directly load their images from the legitimate website.

This becomes an even better criteria, when the image that is loaded is also used as a link. In those cases, legitimate e-mails almost never use an image from one source domain but link to another domain. Thus such discrepancy can be used when trying to classify an e-mail as phishing.

**Pictures loaded from IP addresses:** Last but not least, just as with links, legitimate e-mails do rarely load images from IP addresses instead of registered domains. Thus such a source for an image does nearly always identify a mail as a phishing mail, making this criteria an extremely potent filter criteria.

Looking at these characteristics, it becomes obvious that matching only one criteria does not have to mean that an e-mail is a phishing e-mail, but it becomes more likely with every matching characteristic. Nevertheless some phishing e-mails do not contain any of the above mentioned criteria. Therefore, it is impossible to reliably detect all phishing messages automatically.

Thus both defensive strategies, even combined, cannot provide a complete protection against phishing attacks. They can only reduce the probability of being successfully attacked.

Addressing the problem, that even with filtering and training people fall victim to phishing attacks, many browser vendors, such as Google and Microsoft have created blacklists for phishing websites. These blacklists are then used by the browser to check whether a website is known to be a phishing site and warns the user before he or she can access the site.

However, all these blacklists need to be informed about a phishing site and validate these reports, before they can put it on the blacklist. Thus they need users to report phishing sites to them as fast as possible, such that they can validate them faster and prevent individuals from falling for them. In a

study about the effectiveness of the blacklisting approach [46], the researchers evaluated the blacklisting rate of newly discovered phishing sites which were validated by phishtank [53]. They discovered that Google's blacklist contained around 90% of all live phishing sites they could test, making it a valuable tool to defend users from phishing attempts.
The problem with phishing e-mails that contain attached malware can be tackled if the anti-virus providers are able to provide updated malware signatures faster than people fall for phishing mails spreading the malware. In order to match the speed of new malware and its distribution through phishing mails, anti-virus providers have to rely on malware samples that are reported to them.
The combination of the browser vendors and anti-virus providers efforts against phishing sites and malware provide quite a solid defense as long as new phishing e-mails, phishing sites and malware samples are reported to the designated authorities fast enough.
However, organizations have further options to defend themselves against phishing attacks. The first and most obvious defensive technique are warning e-mails to their members about phishing campaigns. This increases the awareness for the phishing mails and decreases the likelihood that any member of the organization falls for a phishing mail.
Additionally, if the organization knows about the phishing campaign it can also blacklist the phishing website themselves, ensuring that no security breach can occur from that moment on through this particular phishing attack. If the phishing mail distributes malware, they can block the access to potential control servers and try to harden their systems against this type of malware.
Last but not least, if a security breach has already taken place and was detected, the organization can minimize the impact and recover faster.
However, all these options rely on the organization knowing about the phishing attack. Thus internal reporting of phishing attacks is very important and valuable for organizations.
As a conclusion, it can be said, that automatic filtering and human based defensive strategies have proven to be not 100% effective and all other defensive measures rely on the phishing mails to be reported first. Thus reporting phishing e-mails of any kind is important for peoples personal as well as organizations security.

## 2.2 Security Incident Reporting

The main goal of organization security efforts is to prevent successful attacks against the organization to avoid damages. Therefore, the focus of those security efforts lies mainly in the detection and closing of vulnerabilities. However, due to the huge variety of threats and attack vectors, not all of them can be prevented or thwarted. Therefore, organizations establish security measures that take effect after a security breach. Such security measures include the recovery of damaged or manipulated data and business processes.
For these corrective measures to be utilized, a security breach needs to be detected first. A security breach is often noticed through a combination of certain security incidents that occurred in the system. A security incident in this thesis is an event which might compromise an organization. This definition is closely aligned to the definition in RFC 2350 [9]. However, the definition uses the term of compromised

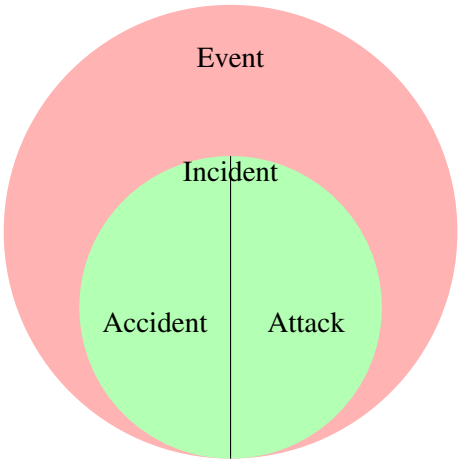


Figure 2.5: Event Space

organizations, which has not been defined yet. An organization has been compromised, if one or more of the following events occurred. The events are that the organization lost sensitive information, or an unauthorized individual gained access to confidential information. Further events are the unauthorized manipulation of data, thereby damaging the data integrity, as well as the denial of access to some data or service through a denial of service attack. Last but not least the misuse or abuse of services as well as damaging some organization system also constitute compromises.
However, also private users can report incidents. In their case the compromised systems would be their personal computer or property and the compromised organization would only be themselves privately. For example many malware attacks target private end users. If an end user notices such an attack, he or she can report this attack to an anti-virus provider or other services.

### 2.2.1 The Reporting Process

In general, security incidents are first noticed by the organizations services. Such a service could for example be the organizations intranet which notices that a user entered a wrong password and username combination. The system itself cannot infer at this point whether the wrong combination came from a typing error, a user who has forgotten his or her password, or through an attacker who tries to brute force a username password combination. Therefore, the best thing to do for these services is to report the incident and logging it. These logs can then be aggregated or directly evaluated by designated personal or automated services which then can infer whether or not the accumulated incidents point to an attack or an accident. Both cases are events and incidents. This can be seen in the event space, see Figure 2.5. If the content of a single service's logs are not enough, then the interesting incidents can be reported further to an authority or designated personal, which might combine the reports of multiple services to gain knowledge about more complex attacks. At the end of the reporting chain should be a computer security incident response team (CSIRT) which should evaluate all reports, identify incidents which lead to compromises and manage the response to successful and not successful attacks.
The response to an attack includes the continuity management, which is tasked with organizing the recovery from the damages created by the attack, as well as preventing further attacks which might

use similar vulnerabilities. Thus the CSIRT is tasked with analyzing the incident chain and finding the original attack source as well as the vulnerability that has been used and removing it. If the vulnerability lies completely under the organizations control, this is quite easy. However, the CSIRTs are well connected and organized. Therefore, CSIRTs can cooperate and analyze incidents outside of the organizations influence. As an example, lets assume that an organization was hit by a DDOS, distributed denial of service, attack. In this case the attack was conducted via a large number of third party systems which were either misused or even taken over by an attacker to attack the organization. The organization itself has relatively few options to prevent a new attack from the same attacker. The most effective strategy would be to reduce the amount of third party systems an attacker can use. Therefore, the CSIRTs identify the participating systems and notify the systems owners about their involvement. This is called network abuse reporting. Moreover, since nearly all attacks are performed over the internet organizations themselves do only possess a very limited influence on the attacking systems. However, the purpose of CSIRTs is to reduce the number of attacks at all. Whereby each reporting organization can reduce the attacks against themselves as well.

### 2.2.2 Network Abuse Reporting

Through network abuse reporting organizations are informed about abusive usage of their internet resources. For example an internet service provide (ISP) will rent out their IP addresses and hosting providers rent out servers to their customers. Thus any customer who misbehaves by abusing the IP address to perform attacks on organizations will always be traced back to the ISP or the hosting provider by the attacked organization. Only the resource owner can identify the abusive user. Furthermore, some users themselves might have been compromised and their machines used by an attacker which happens a lot in real life scenarios. An example for such a takeover operation can be seen in the Mirai bot-net creation from internet of things (IOT) devices.

These victims often remain unaware of the successful attack. For example, the owner of a server does not always notice that somewhere on his or her server a phishing site is hosted, or that apart from handling his or her website and e-mail traffic his or her server is also distributing spam or malware. Nevertheless, those activities are often recognized by external parties, who in turn can report this behavior to the ISP, which in turn can notify the owner of the server or deactivate it completely.

Even if the owner of the server is purposely malicious, reporting the behavior to the ISP is still useful as the ISP can cancel their business deals with the malicious user and thereby remove the malicious server from the web. The reasons why ISPs cooperate are to avoid having their IP addresses blacklisted as well as their reputation damaged.

Organization internally all reporting parties could agree on their own format and processes for reports, such as directly evaluating the logs. But when reporting to an outside authority such as an ISP or an outside CSIRT, the formats and processes have to be agreed upon mutually. Nevertheless, there is no general process specifying how to report network abuse.

In the paper ,*"Why wasn't I notified" Information Security Incident Reporting Demystified* [43], Erka Koivunen describes a theoretical linear model for abuse reporting. This model is visualized in Fig-

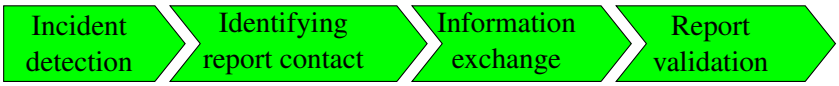


Figure 2.6: abuse reporting process

ure 2.6 and consists of four stages. The first stage is the incident detection stage. In this stage the person or organization which wants to file a network abuse report has to detect the security incident. This is usually done as described in Section 2.2.1.

After an incident has been discovered, the discoverer who wants to report the incident has to find a suitable contact to report to. RFC 2142 [19] made it compulsory to establish a report address of the form *abuse@domain*, but the RFC is often not implemented. Another contact address can be found alongside the domain registration in the whois entries. However, these contacts are often unreliable. Other projects such as the abuse contact register at the regional internet registers (RIR) are still in progress and therefore not complete. Therefore, it is not always possible for reporters to identify the correct contacts.

However, some CSIRTs and clearing-houses accept reports from volunteering individuals and organizations and try to relay those reports to the correct respondent. This sometimes involves relaying it to other CSIRTs or clearinghouses closer to the final receiver. Closer to the final receiver means in this case that the next clearing house is supervising the specific domain. Therefore, the CSIRT has better chances to find a functional abuse contact than the organization and the organization does not need to invest the effort to find an appropriate contact.

The third step is the actual information exchange. For an efficient data exchange it is helpful, if a report format is used that can be read by both parties easily. Therefore, a variety of different reporting formats have been created. The usage and spread of report formats will be discussed later in Section 2.3. Independent of the actual reporting format, every report should describe the incident as well as give evidence for its occurrence.

Last but not least, the report that was received has to be validated. This is necessary as bogus reports can cause damage to the owner of a reported incident source. To validate that the incident has actually occurred, the report receiver can use the information included in the report, to reconstruct the incident that was reported from the report and the local logs. This is done to validate the report and reporter as trustworthy. It has to be noted that the validation process can be shortened and the required proof for a valid report lowered, if the reporter is known to be trustworthy. Therefore, an individual who reports an incident to a large organization has to file a detailed report in order for the report to be accepted as a valid report. However, sometimes it might not be possible to provide sufficient proof for a report, leaving the recipient in the dilemma whether to trust the report or not. This situation can be made easier if the reporter has a trustworthy history with the report receiver. Such a trust relationship is very hard to build up for an individual since it is unlikely that any particular individual will be the incident reporter for multiple incidents from a single source.

Therefore, clearing-houses which accept reports for all incident sources and relay the reports to the

final destination have a distinct advantage. An individual reporter can report all incidents he or she notices to the same clearing house. Trustworthy reporters will provide enough information to prove the occurrence of an incident in most of their reports. Thus a clearing house should after some time be quite certain that reports from such a trustworthy reporter can be trusted, even if they contain quite little proof of an incident. At the same time the clearing house is the report sender for the final report receiver. Hence the clearing house will send more reports to the receiving organization as any individual ever could. Therefore, it is easier for the clearing house to establish a trusted relationship with the organization than for any individual. Thus a report with hardly any evidence coming from a trusted sender to the clearing house is likely to be excepted. The same report that is then sent to the incident source by the clearing house and will also be trusted as the clearing house is a trusted reporter from the organization's point of view. Therefore, the report is more likely to be trusted than if it had been sent by the reporter directly to the incident source. The clearing house in this case functions as a Trusted Third Party and establishes a trust chain between a reporter and a report receiver who might not know each other.

Additionally the clearing house is able to remediate another trust issue. Reporters might fear to reveal personal information to an attacker if they include too much information in the report or even by reporting the incident at all. Sensitive information such as e-mail addresses and IP addresses, could be gathered by a malicious user who received an abuse report to launch further attacks or modify their current attack to be more effective. The clearing house to which an individual or organization reports however can be chosen and thereby should be trusted to not be malicious. Thus the information will not be abused there. The clearing house can then anonymize the reports before sending them to the final recipients. Due to their better trust history, anonymized reports from such clearing houses are still likely to be trusted. At the same time individuals either lack the ability to anonymize the reports, for example they have to send the report form somewhere thereby exposing an e-mail or IP address, or they make their reports less trustworthy by anonymizing them.

### 2.2.3 Efficiency of Security Incident Reporting

The reporting of network abuse can be time and resource consuming. Therefore, the expected outcome of such an report should match the effort. Several studies have been conducted on the efficiency of abuse reports. In his study of security incident reporting behavior, Erka Koivunen [43] describes the effect of six real live scenarios. In five of the six cases, the reports lead to the remediation of the incident source. However, in one case the offender could be caught because of the report, but the DDOS attack continued. Nevertheless, all of these scenarios featured a local CSIRT which sent the reports. Due to the reputation those CSIRTs naturally possess, these results are hardly representative for any reporter.

In another study, performed by Napa, Rafique and Caballero [51], the lifespan of exploit servers were studied. The researches found, that out of the whole 500 servers they were able to track, 13% lived for less than an hour and 60% for less than a day with a median lifetime of 16 hours. However, they also observed that 10% lived longer than a week and 5% living even longer than two weeks. Some servers

could even live up to 2.5 months. In order to measure the influence of abuse reports on these lifespans, the researchers issued reports for 19 long lived exploit servers to the respective abuse contact. Then the time the servers remained active after the report were measured. They found, that these reports proved to be quite ineffective. 61% of their reports remained unanswered. Only seven reports were acknowledged. The median lifetime of all reported servers was 4.3 days. The servers hosted at ISPs who did not acknowledge the report however lived for a median time of 5 days, while servers hosted at ISPs which acknowledged reports only lived for a median time of 3 days. Hence implying that the chance of action being taken after a report is a lot higher for ISPs which acknowledge reports and thereby show interest in receiving these reports. It has to be noted, that all of these lifespans are lower bounds as the researchers can not know whether or not they were the first reporter for any of the reported exploit server. The researchers did analyze the reports for which they had received an acknowledgement further and found that of the seven report receiving ISPs one disconnected the server immediately and another one told the researchers, that they notified the customer. After one day and no action of the customer the researchers filed another report and the ISP did also disconnect the server. Three of the other received acknowledgments promised further information about the actions that would be taken but did not produce any further reply. The servers in those cases lived between two more hours and 1.7 days.

One of the remaining reports was acknowledged by a larger ISP. They responded with the statement that they would take reports seriously, but could not investigate and respond to everyone. The server lived for four more days in this case.

The last case was especially difficult, in this case the ISP responded to the researchers that due to a Dutch regulation, they would need to contact their customer directly and only if the customer would not respond and act in a 5 day time frame, they could contact the ISP again. The researchers complied with the request and the customer did not respond or act within 5 days. Contacting the ISP again resulted in another request to contact the customer. Again the researchers complied, involving the ISP directly this time. However, the customer again did not cooperate, which lead to the ISP requesting the customer to disconnect the server. Thus the server remained alive for more than 5 days after the report. The researchers could however not find any evidence that the 5 day waiting period is part of the Dutch regulations regarding server take-downs.

The researchers draw the conclusion, that ISPs forward the reports to their customers, but do not follow up on abuse reports. Therefore, if a reporter wants to know about the result of their reports and in some case for their report to yield results, they have to follow up on the reports themselves. Additionally this shows the difficulties reporters face when filing reports and the need for unified code of conducts related to the abuse reporting procedures.

Furthermore, comparing the results of both reporting observations, it could be deduced that clearing houses, which look out for their reputation and thereby function as trusted partners can increase the reporting effectiveness by a lot. Especially due to their routine in handling abuse reports.

This was also observed by Hutchings, Clayton and Anderson in their study [39]. They interviewed 24 people, who were actively partaking in website takedowns using abuse reports, about the proce-

dures and results of their efforts. Among those people were experts form law enforcement, companies specialized in taking down fraudulent websites, as well as employees of organizations that were targeted. They found that the amount of reports and takedown operations that are performed regularly directly effect the success rate, as well as the time it takes to have the website taken down. It was especially found that law enforcement officers, who generally perform only very few takedowns each year, are particularly ineffective. The reason for this is assumed to be the lack of routine, as well as the inflexibility of law enforcement operations coupled with the fact that law enforcement often relies on court orders or using their local law enforcement power. Both these concepts are time consuming and not very effective in a globalized world. In contrast specialized organizations have the routine and know how to approach the individual website hosting providers. For example they know who accepts reports to a specific e-mail address, or who wants the reporter to call a call center. Therefore, website takedowns performed by these specialized organizations are the most effective.

### 2.2.4 E-mail Abuse Reporting

E-mail abuse reporting is part of network abuse reporting. However, all attacks that are reported are conducted purely via e-mail. This leads to a few singularities in regard to other network abuse reports. Most network abuse reports are filed by security departments of organizations, researchers or CSIRTs. The reason for this is that these groups possess the expertise as well as access to the important information for detecting those incident. For example the end user in an organization as well as the private end user usually do not know where they would see if someone would try to break into their system. The security department as well as administrators, however should be well aware where to locate access logs, firewall logs and so on, as it is their task to identify potential threats. Thus these groups are especially interested in finding and reporting network abuse directed at their sphere of influence.
However, e-mail abuse is generally not directed at these reporters, but rather at security unaware end users. Additionally, since e-mails can contain private communication, they are not allowed to be read proactively by security experts.
Therefore, to report all e-mail related network abuse, an organization requires the cooperation of all its members. Furthermore, private individuals have to report the e-mails themselves as well. In most network abuse reports, the reporter is well aware, which information can be important for proving the validity of an report as well as which information might be sensitive for himself or the organization. In e-mail abuse reports the receivers of an e-mail report the abusive e-mails, which are largely security unaware end users. These individuals largely lack the knowledge required to report the incident reliably, as well as anonymize reports correctly if needed. Hence the quality of e-mail abuse reports will vary quite a lot.
Moreover, usually the main goal of network abuse reports are to stop an abusive behavior after detection. However, e-mail abuse reports are also important to aid organizations in detecting the same incident on the organizational level. This is important since the security responsible personal themselves can not reliably detect all e-mail related security incidents, therefore it is difficult for them to start the incident response immediately. Instead they often have to wait until they either get notified by

an end user or they notice a compromise of some system through the incident.
This differences make e-mail abuse reporting more important, as well as more challenging than ordinary incident reporting. It does not help, that some abusive e-mails are utilizing social engineering tactics to influence their victim, making them harder to identify than other forms of attack.

## 2.3 Reporting Formats

In the previous section it was shown, that the success of an incident report largely depends on two things. First and foremost, the report receiver has to be willing to receive abuse reports and secondly the information in the report have to be sufficient to prove the reports validity. Additionally, the report has to contain enough information for further processing by the report receiver.
The first factor for the reports success can not be modified by the reporter. However what information is included in the report as well as how the report is structured can be influenced.
The best way to ensure a high quality of reports as well as a universal automated processing of reports would be a globally unique standard for security incident or network abuse reports. But, there exists no universal format for such reports. Thus individuals and organizations, who wanted to report security incidents, invented their own formats to use for their reports.
Originally, handling these reports was done manually, therefore all of the different formats were not a problem. But this manual processing is quite expensive and does not scale well, thus it is not suited for the current as well as future volumes of abuse reports.
Therefore, formats were developed, to standardize the reports and thus ensure a stable quality as well as enable automatic processing. However, up until the creation of this thesis, no single format was able to become the definitive standard for all abuse reports. Nevertheless, a chosen sample of formats will be introduced here.
The first format is the CISL format, which stands for *Common Intrusion Specification Language*. It was developed by the US government Defense Advance Research Project Agency (DARPA) to be used in their *Common Intrusion Detection Framework* and has been published in 1999. The *Common Intrusion Detection Framework* was designed to allow different manufacturers to share information related to security intrusions [70]. CISL is incorporating a LISP like list based data structure, extended by verbs and nouns describing security incidents. This makes CISL very powerful since no limitation for the shared information exists. However, CISL focuses on machine to machine communictation and lies dormant at the moment [72].
Another report format is the *Incident Object Description Exchange Format* (IODEF)[20], which was published in 2007. This format is described in the RFC 5070 and focuses on increasing the automation capabilities when processing incident reports. However, just as CISL it focuses on intrusion detection incidents. The structure of IODEF is based on the XML language.
The next format is the CAIF format, which stands for *Common Announcement Interchange Format* and was published in 2002. It was developed by the Stuttgart University RUSCERT as an XML-based format for exchanging security announcements [61]. CAIF allows for information about vulnerabilities, incidents and problems as well as solutions to be included in a single report. Furthermore, multiple

unrelated incidents can be reported in the same report as well. CAIF is also designed for human to human communication.
A slightly different approach was taken with the ARF format. This format was created by Yakov Shafranovich, as a format to report spam messages. This draft was then taken up by the *Message Anti Abuse Working Group* which formed the *Message Abuse Reporting Group* (MARF) to finalize the format. In 2010 this group published an updated standard ARF [65]. An ARF report consists of three MIME parts. The first MIME part is a human readable description of the report. In the second MIME part is a machine readable description of the report, which has the content type "message/feedback-report". In the third MIME part either the complete e-mail or a complete copy of the e-mail headers is attached. Through the three MIME parts the ARF standard is able to support machine to human communication. However, it still remains limited to reporting spam.
The last reporting format which is going to be presented here is the X-ARF format.

### 2.3.1 X-ARF

X-ARF is an extension to the abuse reporting format [65]. It was created to extend the types of incidents that could be reported with the previously known format, which did only allow for spam e-mails to be reported.
In order to acquire the flexibility necessary for reporting a variety of different incident types, X-ARF is designed as a very flexible format. Therefore X-ARF only defines the structure of the report as well as some key information, which have to be included. The information that the X-ARF standard requires all reports to contain is chosen to represent the bare minimum in order to validate a report [1]. These minimalistic information contain the source of the incident, the time at which the incident occurred, the reporters e-mail address and a unique id that can be used to refer back to this report. The flexibility of X-ARF comes from the support of report schemata. The schemata specify which further information is contained in the individual report. Each schema usually defines one incident type. This allows the schema to specify only information that are relevant to the reported incident to be included in the report itself. Thus removing overhead in the reports and at the same time allow the format to be adapted to new incidents rather easily.
Reports in the X-ARF format are sent via e-mail. Therefore, these reports are structured using the MIME standard. In the first version of the X-ARF standard, version 0.1, each X-ARF report contained an *X-ARF* header field which was set to true. However, in the second iteration, the header, which identifies X-ARF reports changed, since *X-* is reserved for the X-e-mail headers. Therefore, as of version 0.2 an *X-XARF* header field is used. This new header can also differentiate between three different report types. The three message types are *PLAIN* , *BULK* and *SECURE*. *BULK* and *SECURE* X-ARF reports were introduced in 2011, by Sven Übelacker. Generally X-ARF messages are send in the *PLAIN* format, however in certain situations one of the other two types might be better suited for the task. Additionally the field *Auto-Submitted* should be present and be set to the value *auto-generated*. This is necessary in order to comply with RFC 3834. Last but not least, the *Content-Type* field has to be set corresponding to the X-ARF message type as well.

### Plain X-ARF messages

Plain X-ARF messages are the most common type for X-ARF reports. The reason for this is that they are the easiest to generate and validate. They are used to report individual incidents unencrypted and unsigned. The *X-XARF* field value identifying this type is *PLAIN* and the content type of the e-mail has to be set to *multipart/mixed*. Plain X-ARF messages consist of three parts. These threes part are also represented in the message body as three different MIME parts.
The first part of such a report is a human-readable description of the reported incident. It should be written in full sentences, preferably in the English language, and contain the basic information about the incident. Thus this part should include information about the type of incident that is reported, when the incident occurred or, if the time at which the incident occurred is unknown, when the incident was noticed. Furthermore, it should also include information about the source of the incident. At the end of the message a short notification should alert the recipient that the X-ARF format was used and where the recipient can obtain information about the format and how to parse the information in the report correctly.
The second part of the report contains the information specified by the container as well as the X-ARF format itself. It serves as an information pool for the recipient, such that the recipient can use this information to gain a deeper understanding about the incident that was reported. This information is delivered as the YAML representation of a JSON object.

**YAML:** YAML is a human readable data serialization standard [24]. It is designed to provide an easy readable, programming language independent way to serialize data and can be used in a variety of contexts. The data serialization is based on key value pairs, just as JSON. However, it removes the brackets and other formatting issues which make JSON and XML hard to read for humans. To ensure easy usage in a large number of programming languages it provides support for the typical data types in agile programming languages. For further information and the syntactical information please refer to the YAML specification [5].

The YAML representation consists solely of key value pairs which have to match the specifications in the X-ARF report container schema. The report schemata in X-ARF have to contain the fields required by the X-ARF standard, but on top of these fields they can define additional information, specific to the report type the individual schema describes. The schema is written in JSON, thereby allowing the recipient of a report to validate the information parsed from the YAML part against the schemata. This allows the recipient to spot broken reports before trying to process them. Additionally, if the schema definition is tight enough this also limits the possibilities of an attacker to create malicious reports. An example definition of such a schema is given in Listing 2.1. In this schema only the necessary fields defined in the X-ARF standard are included for readability reasons. An example for a report which fulfills the schema is given in the Listing 2.2.

The mandatory fields for any X-ARF report schemata are *Reported-From*, *Category*, *Report-Type*, *User-Agent*, *Report-ID*, *Date*, *Source*, *Source-Type*, *Attachment* and *Schema-URL*. The usage of these

Listing 2.1: Example X-ARF Report Schemata

```
{"description":"An Example Report Schema",
 "type":"object",
 "properties":{
   "Reported-From":{
     "type":"string",
     "format":"email"
   },
   "Report-ID":{
     "type":"string",
     "format":"email"
   },
   "Category":{
     "type":"string",
     "enum":["fraud"]
   },
   "Report-Type":{
     "type":"string",
     "enum":["example_xarf_report"]
   },
   "Service":{
     "type":"string"
   },
  "Port":{
     "type":"integer"
   },
   "User-Agent":{
     "type":"string"
   },
   "Date":{
     "type":"string",
     "format":"date-time"
   },
   "Source":{
     "type":"string"
   },
   "Source-Type":{
     "type":"string",
     "enum":["ipv4","ipv6","ip-address","uri"]
   },
   "Attachment":{
                 "type":"string",
                 "enum":["None", "text/plain"]
   },
   "Schema-URL":{
     "type":"string",
     "format":"uri"
   }
 }
}
```

Listing 2.2: Example X-ARF Report Based on the Example Schema

```
Reported-From: reporter@organization.com
Report-ID: 1234@example.org
Category: info
Report-Type: example_xarf_report
Service: http
Port: 80
User-Agent: ExampleUserAgent V1
Date: Mon, 10 Feb 2017 22:32:19 +0200
Source: http://domain.tld/images/login.html
Source-Type: uri
Attachment: text/plain
Schema-URL: http://www.example.org/schema/example.json
```

fields will now be looked at individually.

**Reported-From:** The Reported-From field contains the e-mail address of the reporter. This address can therefore be used to direct questions and feedback regarding the report back to the reporting party. An example for value for this field would be could be an address like *Reported-From: reporter@organization.com* seen in Listing 2.2.

**Category:** This field is used to store a short constant describing the type of report. The type of report is chosen depending on the reported incident as well as the reporters own categorization. It is possible for X-ARF schemata to restrict the possible report types for a report. The constant values from which one has to be chosen are:

- abuse : For any attack reports such as login attacks or DDOS attempts.
- fraud : For abuse related to fraudulent behavior such as credit card fraud.
- auth : For abuse or failures of authentication methods.
- info : For purely informational reports such as blacklistings.
- private: For sensitive information shared only between two parties.

**Report-Type:** This Report-Type describes the specific type of the report. The type of a report is always a string, which is defined in the report schemata that is used. This string is usually used to also identify the used schemata. Therefore, it is uniquely defined for each report schema. To prevent doubling a report type value, these unique identifiers are chosen by the X-ARF community. The schemata should restrict this value to exactly one possible constant to prevent misuse. An example for the definition and usage of this field can be seen in the example Listings 2.2 and 2.1, in which the report type is defined as the string *example_xarf_report*.

**User-Agent:** The User-Agent field provides an identifier for the tool used to create the specific report. It is helpful in identifying tools which create broken reports, as well as tools which might

automatically generate reports but behave unreliably. Furthermore, as stated in RFC 1945 [6], it can be used to tailor the feedback to a report specifically to the sending system.

**Report-ID:** The Report-ID is an important field, as it provides each report with a unique identifier. Thus it enables the report receiver and report sender to exchange information and feedback about specific reports. Therefore, the id has to be unique across domains and time. Thus it is advised to construct the id from a compact domain specific unique identifier, followed by an @ (at) symbol and the reporting domain. As long as the first identifier is unique for the reporting domain and the domain remains consistent over time, the report id will be unique across time and domains. The major advantage of such a generation scheme for unique ids is that it can be enforced in the schemata. For example in the example schemata in Listing 2.1, the report id field is restricted to contain strings in e-mail address format. Thus only unique ids of the described form will be regarded as valid.

**Date:** In the Date field, the date of the incident should be conserved. However, if the date of the incident can not be determined anymore, then the date of the incident discovery should be used instead. The format in which this date should be given should be restricted to the date-time format inside of the spezific report schematas, as seen in the example schemata 2.1. The current X-ARF standard, version 0.2, advices the use of the time format described in RFC 3339 [42]. However, due to compatibility reasons with older versions, the in RFC 2822 defined time format should also be accepted. But should this older format be used, the format restrictions for the Date field should be removed from the schema. Nevertheless, this is for compatibility reasons not required, hence any parser needs to check for both formats, regardless of the presence of the format information.

**Source:** The Source field is used to state the source of the incident. The form in which the source is stated is quite flexible. For example in the Listing 2.2 an URL is used as source specification. In other cases however this specification can also be done by stating IP or e-mail addresses. Thus the source field is simply defined as any string. The validity of the Source field can however be checked by looking at the Source-Type field, which is also mandatory in X-ARF reports.

**Source-Type:** The Source-Type field specifies the type of source specification, given in the Source field. The potential source types are restricted to a small set of allowed types:

**ipv4:** As the name suggests, this type is used to specify that a standard four byte IP address, compliant with RFC 791, is given as the incidents source, such as *192.0.2.1*.

**ip-address:** This type is another identifier which specifies that a RFC 791 compliant IP address is used.

**ipv6:** The ipv6 identifier specifies, that a RFC 2460 compliant IP address is given in the source field.

**uri:** This identifier is used to specify that a resource links such as *www.malware.evil/mal* is used as an incident source. Such an incident source is especially useful in malware reports.

**domain:** The domain specifier can be used to specify a whole domain as malicious.

**email:** The last supported source type is the e-mail specifier which determines that an e-mail address is contained in the source field.

**Attachment:** In the Attachment field the content of the third X-ARF report part is specified. Using this field, it can also be specified, that no third part exists. This is however not advised as it leads to proof for the incident beeing missing. Furthermore, the attachment field also specifies what kind of attachment is contained, thereby allowing for an at least partly automatic interpretation of the attachment. The type of attachment is specified by the corresponding MIME type.

**Schema-URL:** Last but not least, the Schema-URL is one of the most important fields. It specifies a url under which the JSON schema of the report can be found. Therefore, a receiver of a report can use this URL to look at the schema and verify the reports correctness. Furthermore, only inside of the individual schema all fields included in the report are described. Therefore, without the schema, the receiver might not be able to gain all information available in the report.

Besides the mandatory fields, which have to be included in every X-ARF report, X-ARF specifies a few optional fields. These optional fields do not have to be included, but it is advised to include them in the reports. These fields are *Version*, *Occurrences* and *TLP*.

**Version:** The version field contains the version of the X-ARF standard that is used to generate the report. This ensures that reports coming from old reporting tools using older versions of the standard can be detected and potentially still be validated and parsed.

**Occurrences:** This field is used to specify how many incidents of the same type have occurred and are reported in this report. This field is optional as each incident should be reported individually, leaving the occurrences field at the value one. However, for some incidents, the malicious nature becomes only apparent after a certain amount of occurrences. For example in a login attack a single failed authentication does not reliably identify an attack. In this case multiple occurrences are necessary to identify an attack.

**TLP:** TLP stands for Traffic Light Protocol, which is short for Information Sharing Traffic Light Protocol [74]. It is used to specify the sensitivity of the information shared in the report. The TLP field contains one of the four possible values red, amber, green, white. Each of these values defines a different confidentiality level:

- **red**

  Red specifies the highest confidentiality level. It states that all the information in the communicative exchange can only be shared between the communicating parties. Therefore, the report cannot be shared with other authorities. Even more specific, no details about

the information exchange can be shared with any individual not directly involved in the communication.

- **amber**
  The amber confidentiality option allows the information to be shared with members of the organizations, the communicating parties belong to, if these members were either directly involved in the communication or if they *absolutely need to know* in order to take action.
- **green**
  If the green level is used, then information can be shared with other organizations or departments inside of the organizations the communicating parties belong to. However, the information can only be shared with either the network security or information assurance departments. Furthermore, it is strictly forbidden to publish the information or post it online.
- **white**
  The white confidentiality level is given to information that is open to the public. As long as the copyright is not infringed, the information can be published and shared freely by and with anybody. The copyright in this case refers to the copyright a creator has over his documents. Regarding reports this means that if a copyrighted work is reported with the TLP white, it is not allowed to distribute the report with the copyrighted material without the explicit permission of the copyright owner, as it would infringe the copyright.

The third part of an X-ARF report contains the evidence for the incident. Such evidence can be log entries, captured internet traffic or similar documents. Most often the documents contain sensitive information of the reporting party or other unrelated network users. Thus the documents used as evidence can be cropped and anoymized in order to protect the sensitive information.
Combining the report parts results in a report structure as seen in Figure 2.7.

### Secure X-ARF messages

Secure X-ARF messages can be identified by the X-XARF field being set to SECURE. This X-ARF report type was introduced in the version 0.2 of the standard. They can be used to send signed and/or encrypted reports. Therefore, the content type of the report has to be set to one of *multipart/signed*, *multipart/encrypted*, or *application/pkcs7-mime*. When using the secure X-ARF messages, a plain X-ARF message is created. This plain message then constitutes a complete RFC 2822 conform plain X-ARF report. This report is then used as a container, which can be signed and/or encrypted and thereby create the secure X-ARF message.

### Bulk X-ARF messages

Bulk X-ARF messages are used to report multiple incidents economically in a single report. An X-ARF report is of the bulk type if the X-XARF field is set to BULK. This report type has a content type of *multipart/mixed*, just as the plain X-ARF reports do. However, they do not contain the typical three MIME parts as the plain message does, rather the bulk message body is made up of other plain

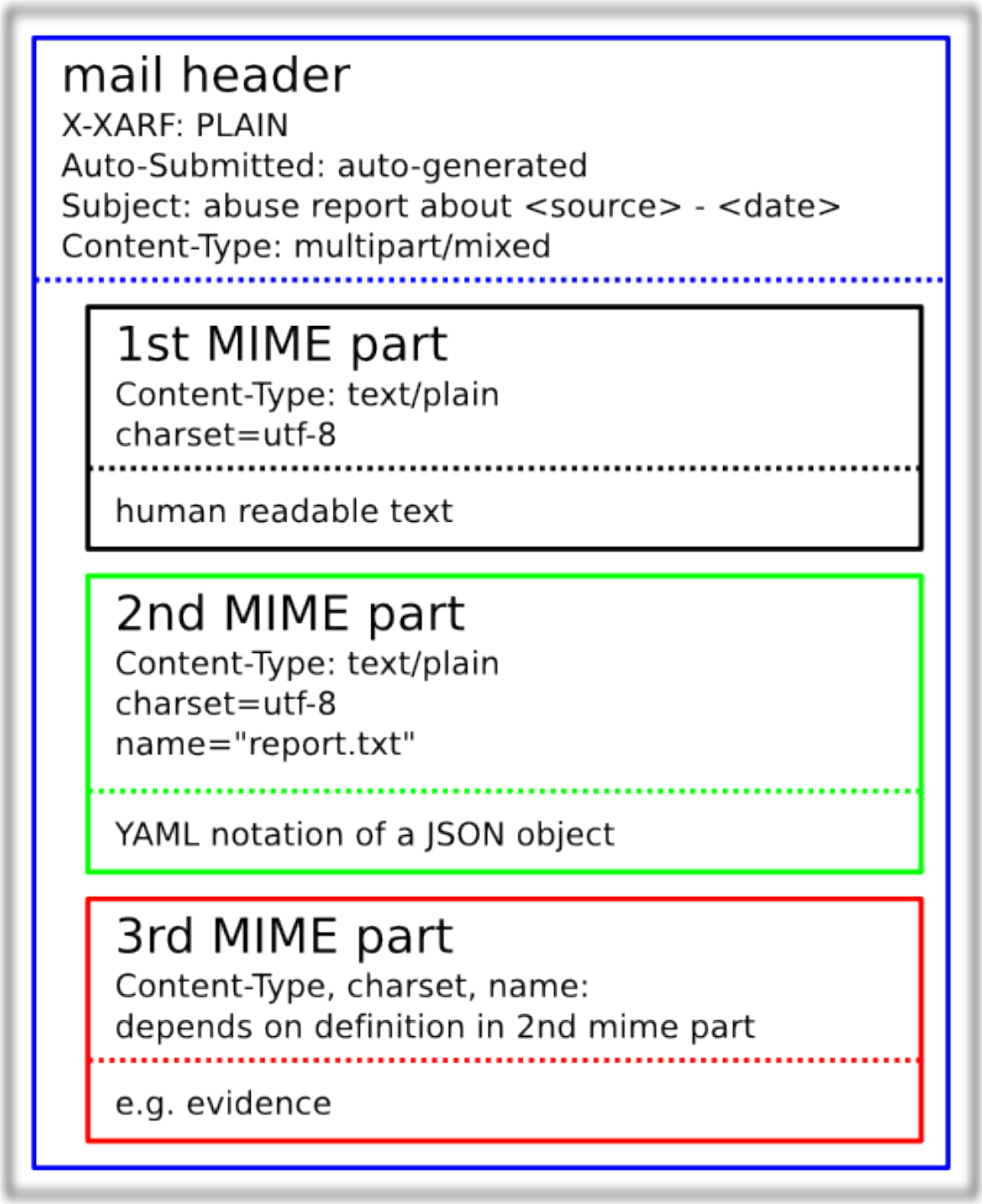


Figure 2.7: Schematic X-ARF Report [79]

and secure X-ARF reports. The number of reports which can be combined into a single bulk report is unlimited. To interpret these bulk messages, a receiver can just iterate over all contained messages and evaluate them individually, as each message contained in the bulk X-ARF report is a complete valid report on its own. This gives the main advantage, that the receivers of bulk reports can simply unpack all of the reports and handle each report on their own. Additionally, clearing houses can simply gather all reports, pack them into a bulk report and send this report to an abuse contact. This makes the usage of bulk messages very easy.

# 3 Survey Design

In order to gather information about the susceptibility of end users regarding phishing e-mails as well as their knowledge regarding reporting, a survey was performed. The survey was conducted completely anonymous to minimize the effects of social desirability. It consisted of four parts, each of these parts is supposed to gather information related to a different field. The first part gathers some information about the participants themselves. These information are related solely to the personality traits, the risk perception and taking as well as the general computer security behavior. In the second part a small scale phishing experiment was conducted, while in the third section the participant's experience related to phishing e-mails is evaluated. The fourth question group is gathering information about the participant's knowledge and behavior regarding security incident reporting. The last section is used to gather information about the participant's desires and requirements for an automated reporting tool.

## 3.1 Population Sample

The survey was conducted with a group of employees in a small company. All employees worked for the same company, which develops individual business software as well as provides IT consulting services for their clients. The company has a total of 80 employees, however only 42 volunteered to participate in the survey. The participants have an IT background and can be regarded as tech savvy and more knowledgeable regarding IT security practices than the average end user. However, none of the participants has an explicit IT security background other than a short introduction received in their university studies.

Since participation was completely anonymous, no information about age, gender and work experience were gathered as those information could potentially identify the participant. This, would have increased problems of social desirability. However, two experiments performed by El Zarka, H. Bhojani and Darwisch [50] could not produce conclusive evidence, that gender or age were meaningful factors regarding the susceptibility towards phishing. However, many other papers, such as the study by Sheng, Holbrook, Kumaraguru, Cranor and Downs [67], have proven that these factors indeed make a difference regarding phishing susceptibility. Therefore, the gathered data has to be used cautiously when extrapolating form it.

## 3.2 Personality and Risk Taking Question Group

The personality and risk behavior questionnaire part consists of the Domain-Specific Risk Taking (DOSPERT) scale [7], the Security Behavior Intention Scale (SeBIS) [23], the Barratt Impulsiveness Scale-Brief (BISbrief) [71] and the Ten Item Personality Inventory(TIPI) [33].

This questionnaire section serves two purposes. First and foremost it is used to gather the information

about personality, risk taking and security behavior, to correlate these information with the phishing and incident reporting related information. The second purpose of this section is to familiarize the participants with the survey in general.

### 3.2.1 Domain-Specific Risk Taking Scale

The Domain-Specific Risk Taking scale was designed by Weber, Blais, and Betz in 2002 and refined by Weber and Blais in 2006, to measure the risk taking behavior regarding five different domains (ethical, financial, health/safety, social and recreations). This was necessary as the researches concluded, that a person might be quite willing to take some recreational risks, such as bungee jumping, however, the same person can be very risk averse regarding financial decisions. In this scale, risky behavior is defined as behavior that can lead to dire consequences.

The refined DOSPERT scale consists of 30 questions, six per risk domain, and has been verified for multiple research domains as well as multiple translations [28]. In the survey the English version was chosen, since it is the most used and allows for the questionnaire to remain consistent in the language that is used, as some other scales are only validated in English.

Furthermore, not all risk domains are interesting regarding phishing. For example the attitude of a person towards ethical, recreational or health related risks is not affecting a persons susceptibility to phishing.

The financial risk affinity might be affecting the susceptibility towards phishing. But since other social factors have a much higher impact on the susceptibility to phishing, leaving out the financial risk domain to shorten the survey, appeared to be sensible.

However, the social risk domain describes the potential level of influence the Caldini's six principles, described in Section 2.1.6, have on an individual. Therefore, only the the five questions for this risk taking behavior domain were used in the survey.

### 3.2.2 Security Behavior Intention Scale

The Security Behavior Intention Scale (SeBIS), developed by Egelmann and Peer [23], is designed to measure the security behavior of end users. The scale consists of 16 questions in which the participants state how often they follow the behavior, advised by security experts. An example would be updating their software. These information provide an estimation about the security awareness of the participants as well as a rough estimate of their security related behavior. However, the security behavior should not be trusted upon too much, due to a potential social bias. Social bias stems from the fact, that people might feel ashamed to admit mistakes or socially ill-advised behavior. Egelmann and Peer tried to resolve this problem by titling their survey as a computer behavior survey instead of a security survey. Additionally, they let all participants answer the ten item Strahan-Gerbasi version of the Marlowe-Crowne Social Desirability Scale [75], to evaluate the influence of the social bias. But in this survey the problem regarding social bias could not be solved the same way. First of all, the survey title has to match the purpose, which is to evaluate end users incident reporting knowledge and usability requirements for reporting tools. Titling it otherwise would lead participant to feel deceived,

as the purpose of the survey becomes clear to anyone in the later stages of the survey.
Secondly, performing another ten item questionnaire about the social bias of participants lengthens the survey further without directly benefiting the main purpose. Lengthening the survey would lead to fewer willing participants and thus reduce the significance of the survey.
Nevertheless, to limit the risk of skewed results due to social bias, the complete survey is performed online, without supervision. Therefore, it is anonymous, which reduces the incentive to answer questions according to social norms instead of actual behavior or opinion.

### 3.2.3 Barratt Impulsiveness Scale-Brief

The Barratt's Impulsiveness Scale (BIS) [69, 29] is a very common questionnaire to determine a persons impulsiveness. This personality trait is important to rate the results regarding the phishing susceptibility of a single end user and thereby make valuable predictions about the overall end user susceptibility to phishing. However, the BIS is quite time consuming to perform, as it is a quite large questionnaire with 30 items. Therefore, the Barratt Impulsiveness Scale-Brief (BISbrief) is used which only contains eight items while maintaining the information quality that is obtained by the original scale [29]. This makes the BISbrief a perfect fit for the conducted survey, because the personality evaluation of the participant is only a small part of the overall survey and should not be exhausting to the participants.

### 3.2.4 Ten Item Personality Inventory

The Ten Item Personality Inventory(TIPI) is a personality questionnaire gathering information about the participants regarding the five basic personality tendencies [48], which was described in Section 2.1.6. Most other questionnaires trying to collect information about personality traits use multiple questions for each of these five tendencies. An example for such a questionnaire is the Big five inventory scale, which consists of 44 items [57]. However, this leads to many similar questions, which can annoy participants.
Therefore, TIPI reduced the number of questions to ten, which are phrased as "*I see myself as:*". These questions correspond to the ten poles of the five tendencies. Thus no two questions appear too similar. Additionally, answering these ten questions is a lot faster than the bigger scales. It takes just one minute to answer the TIPI test. This results in much more cooperation from the participants after answering these questions.
However, shortening the questionnaire comes with a price. While other questionnaires can focus on internal consistency of a participants answers, which is improved by multiple closely related questions, TIPI cannot. The reason for this is that only two questions per tendency are simply not enough to establish the necessary overlap. Therefore, TIPI only focuses on the validity of the questions and answers.
Nevertheless, it has been shown that the results of TIPI and other longer questionnaires do converge [33]. Therefore, TIPI provides a valuable estimate for the five personality tendencies of a participant, while reducing the time consumption and participants annoyance of the other questionnaires. For the survey in this thesis the loss in internal consistency is accepted in order to reduce the time the participants need to complete the survey.

## 3.3 Phishing Experiment

The second part of the survey consists of a small phishing quiz, which was designed by sonicwall [40]. In this quiz the participants are presented with a set of ten e-mails. The task is to identify, for each e-mail individually, whether the e-mail is legitimate or not. The e-mails are all in English and are taken from a real life samples of phishing e-mails. In order to allow users to detect the phishing e-mails, the URLs of links, which are present in the mail, are shown below the e-mail itself. However, the participants cannot interact with the e-mails. At the end of the quiz, the participants are presented with a page showing their performance in the test, as well as the missed hints for false classifications. The phishing quiz does always consist of the same ten mails, from which three are legitimate and seven are phishing e-mails. This fact was unknown to the participants in order to prevent them form tactically classifying e-mail instead of looking at the clues.

Upon completing this task the participants return to the survey and enter their results. Furthermore, they are asked, how confident they were when they were categorizing the e-mails. This allows for more diversity in the results of the phishing quiz and reduces the effect of guessing in the multiple choice test. If a participant was unsure how to classify an e-mail, he or she was then asked whether he or she would have liked to ask a more qualified person to help with the classification. If they answered this question with a “No”, then they were also asked why they would not want to ask for an experts opinion on the e-mail.

## 3.4 Experience with Phishing

In the third section, participants were questioned about their experiences with phishing e-mails. The two questions asked them if they perceive phishing attacks to be a threat for them personally and for an organization they belong to. These questions are used to determine how the participants perceive the risk of falling for phishing attempts.

Afterwards, they were asked if they ever received a phishing e-mail themselves and whether they know someone who fell for a phishing attack. Together with the fourth and fifth question, which asks whether they themselves fell for a phishing attack directed at their private or professional e-mail account, these questions give an overview whether the people participating in this survey or their friends or family fell for phishing attacks in the past and how this effects their desires for reporting phishing attacks.

In all those questions falling for a phishing mail was phrased as *trusted a phishing e-mail*, which was deliberately chosen to avoid the social stigmata regarding being deceived or being the victim of something. The word trusting has a positive connotation, in contrast to the terms *being deceived* or *becoming a victim of* which are generally perceived negatively. Therefore, trusting something is not as socially ill-advised as being deceived would be and thereby reduces the risk of a social bias. Furthermore, the order of the question is also chosen to alleviate the potential social stigmata related to being a victim of phishing scheme, as the first question ask if they know someone (this could be the participant himself) who fell for a phishing mail. Therefore, if they answer the two questions afterwards truthfully, they will not be the only victims and therefore not the only one who fell for such a scam. However, social bias cannot be ruled out for these questions.

After the questions about falling for phishing attacks, the participants are asked who should be responsible to prevent successful phishing attacks. Here they could choose between the options "*every member of the organization themselves*", "*the organization for all the employees through administrators and their security department*" or "*employees and the organization together*".
This question group ends with the question how the participants feel about antivirus software and the protection it provides from phishing attacks as well as their opinion on the usefulness of phishing warnings.

## 3.5 Experience with Incident Reporting

The fourth section is concerned with the participants knowledge and behavior regarding security incident reporting. The first two questions in this group ask whether the participants have ever reported any incident to an authority inside or outside of their organization and whether they have reported an incident which occurred in their private lives to an authority. This is done to determine the participants current reporting behavior and the incident reporting culture inside of the organization.
The distinction between these questions is related to the environment in which the participant reported security incidents. Some organizations have rules regarding the reporting of security incidents. Therefore, end users in these organizations might regularly report security incidents internally, however in their private life they might not do so.
At the end of the question group they were asked, if they know to whom they can report incidents and if they do, to which authority they could report to.

## 3.6 Reporting Tool Wishes

The last section of the questionnaire focused on the opinion and wishes of the participants regarding existence and usability of automated e-mail reporting tools. Therefore, the next two questions asked the participants whether they even would want to report such e-mails. The first question focused on their private e-mail account, while the second question focused on the desire to report e-mails directed at their professional account. Then they are asked whether they want to receive feedback and in which form they would like to receive feedback. The last question in this group as well as the survey was how fast the participants would like to receive feedback to their reports.

## 3.7 Survey Results

From the sample group 42 people participated in the survey, however only 28 people completed the whole survey. The other 14 people did only answer the first question group and were therefore excluded from the analysis. Analyzing the BIS scale showed that the participants were not very impulsive, as they reached an average value of 1.83 on the BIS, which ranges from 1 to 4. The standard deviation was 0.38, with a minimum of 1 and a maximum of 2.88. However, their social risk behavior measured by the DOSPERT-RT scale showed, that they are quite willing to take risks in that domain. The average value for DOSPERT lays at 4.88 with the standard deviation of 0.88 on a seven point likert scale. This observation matches the results the TIPI provided, in which the participants scored exceptionally high

Table 3.1: Personality Trait Scores According to TIPI

| Trait | openness | conscientious | stability | agreeableness | extraversion |
|---|---|---|---|---|---|
| AVG | 5.4 | 5.75 | 5.1 | 4.64 | 3.45 |
| SD | 1 | 1.17 | 1.12 | 1.07 | 1.71 |

values in regarding the openness to new experiences. They scored an average of 5.4 and a standard deviation of 1 on the same seven point likert scale. Furthermore, TIPI revealed the participants to be very conscientious and emotionally stable. On both traits the participants on average scored over 5 points, with a conscientiousness scoring of 5.75 and an emotional stability rating of 5.1. The other results of the TIPI can be seen in the Table 3.1.

Due to the small standard deviations, it can be seen that the sample is rather homogeneous regarding their personality traits. They are in general open to new experiences, which comes as no surprise as they work in an software development and IT consulting firm and openness to experience is an indicator for people to be more technologically interested. Additionally they are rather conscientious and emotionally stable. Furthermore, they are willing to take social risk while at the same time not being extremely impulsive. This could stem from the consulting activity of the employer, which requires consultants to not be impulsive but still be willing to take social risks, such as disagreeing with authorities. However, the group does not appear to be so homogeneous regarding their extraversion. The standard deviation of 1.71 defines a fluctuation of the result of about 24% of the possible scores.

Last but not least looking at the results of the SeBIS scores shows, that the participants are scoring a bit above average in the password security (3.74), updating (3.54) as well as security awareness (3.58). The standard deviations are in this case 0.88, 0.71 and 0.74 showing that the security awareness regarding these properties is also quite homogeneously distributed. However, the participants score with 4.3 and a standard deviation of 0.75 significantly higher regarding device securement than average. Nevertheless, it is expected that people employed in software development and IT consulting are a little bit more security aware and follow secure practices a little bit better than the average individual would.

These results indicate that the participants should be doing rather well in the phishing test. The reason for this is that the security awareness paired with the conscientiousness and the clear task to separate the phishing mails from the valid ones should mean that they will observe the potential phishing mails quite closely and therefore find most of them. Analyzing the responses regarding the online phishing test showed that four participants did not understand the questions regarding the phishing test, which resulted in them entering false data. Those four participants were excluded from any analysis of the phishing test results. Those results reveal that the average participant was able to identify 7.4 e-mails correctly and believed 1.2 phishing e-mails to be legitimate. Therefore from the 7 phishing e-mails 17% were able to fool the average participant. This is significantly higher than the results of the experiments which lay between 9 and 12 % for non spear phishing e-mails, described in Section 2.1.5. However the task in the phishing experiment was also much harder than in the other experiments. First and foremost, the participants were unable to interact with the phishing e-mails. Therefore, they were completely relying on the information that were given in the quiz and had to find where the information

were located. Some participants voluntarily contacted the researcher to provide feedback to the phishing experiment. In those statements some participants stated, that they did not find where the location underlying URL of a link was presented.
This also showed in the responses to the question how sure they were when categorizing the e-mails. From the 24 participants whose results were counted in the analysis 22 answered that they were unsure about at least 1 or 2 e-mails. Thus only 2 people were absolutely sure to have chosen correctly. From the group that was uncertain about some e-mails roughly 77% would have liked to ask for assistance when facing e-mails they were unsure about. The other 23% were then asked why they would not ask for assistance, to which the majority answered, that they would have felt sure if they would have been able to hover over the links and interact with the e-mails more freely. However, one person stated, that he or she felt it would take to much time to ask someone for assistance. Another person stated, that he or she would not ask for assistance because he or she would never follow any link in an e-mail he or she was unsure about.
Nevertheless, 28% of the participants stated that they had learned new techniques to identify phishing e-mails during the test and another 17% were unsure whether they learned new detection criteria or not. This shows that even in this tech-savvy group of participants not all deceptive techniques used in these example phishing mails were known and secondly that the participants understood the explanation at the end of the test.
The results of the phishing experiment match the participants phishing risk perception. Asked whether they agreed that phishing e-mails are a risk for their organization, roughly 90% agreed or strongly agreed. Additionally, when asked whether phishing e-mails are a risk for them personally, 78% agreed. These high percentages might come from the experience the participants had with phishing. In the questionnaire about 40% stated that they know someone who fell for a phishing attempt and roughly 10% also stated, that they themselves might have fallen for a phishing e-mail send to them in their private life. However, these two questions are strongly affected by a potential social bias, as falling for a phishing mail can have a social stigmata. Thus the percentages could be to low.
The participants did largely agree that the responsibility for preventing the success of phishing attacks should lie by the employer as well as the employee together. A total of 79% voted for sharing the responsibility, while 14% believed the responsibility should reside by the employee and 7% gave the responsibility to the employer.
When asked if they agree that anti-virus software and firewalls are to prevent successful phishing attacks, only 18% agreed with the statement, but 43% disagreed and 39% even strongly disagreed with the statement. This indicates, that the participants know that phishing e-mails try and are often able to bypass such standard defenses. However, when presented with the statement, that warning e-mails regarding phishing e-mails that other employees did receive help them to identify phishing e-mails themselves, 86% agree with the statement and only 14% disagree or strongly disagree. Therefore, a majority believes, that phishing warnings which can be created following a phishing report can be a valuable tool in preventing phishing attacks.
This is also represented in the desire of the participants to report. In Figure 3.2 and 3.1 the answers to

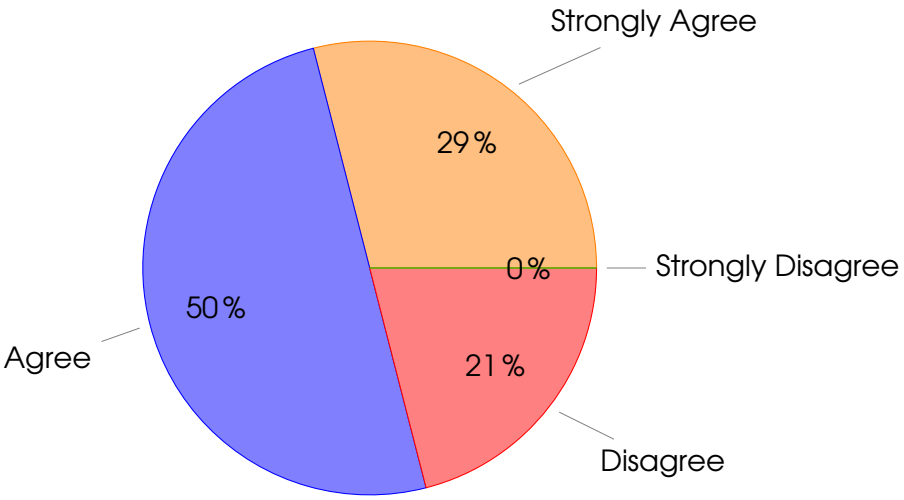


Figure 3.1: Responses to the question: Do you agree with the following statement: ”I want to report suspicious e-mails sent to my private e-mail account to an outside authority to protect other individuals from potential phishing attacks”

Table 3.2: Known authorities with the amount of people who named them as an authority

| Authority | Amount |
|---|---|
| administrators | 5 |
| sec. department | 2 |
| superiors | 2 |
| project team | 1 |

the questions regarding the desire to report suspicious e-mails can be seen. In both cases over 75% of participants were agreeing with the statement, that they would like to report suspicious e-mails. However, it becomes clear, that the participants were far more willing to report e-mails in their professional life as in their private life. These values imply that among the participants incident reporting is quite common. However, in the questionnaire only 29% have reported any security incident in their professional career so far and incidents that occurred in their private life were only reported by 14%. The reason for this chasm can come from the fact that only a quarter of all participants knew any authority to which they could report incidents to. It has to be noted, that among the 29% who did report an incident not all did know an authority to whom they should report to. They reported the incident to the team leader or someone else, but no one with any specific authority. Those who knew an authority were asked to name the authority they knew. This resulted in the following named groups: the administrators, the superiors, the security department and the project team. The distribution of known authorities can be seen in Table 3.2. All these authorities are organization internal entities. Therefore, private reports to these authorities might not be desired. Furthermore, due to the variety of stated authorities and the fact that all participants are employees of the same company, it can be seen that no consistent report processing can be employed if the employees do not have a unique authority to report to.
Until now, it could be seen that the employees want to report suspicious e-mails, however, they do not know exactly to whom they should report them. Thus they might need assistance with the selection of

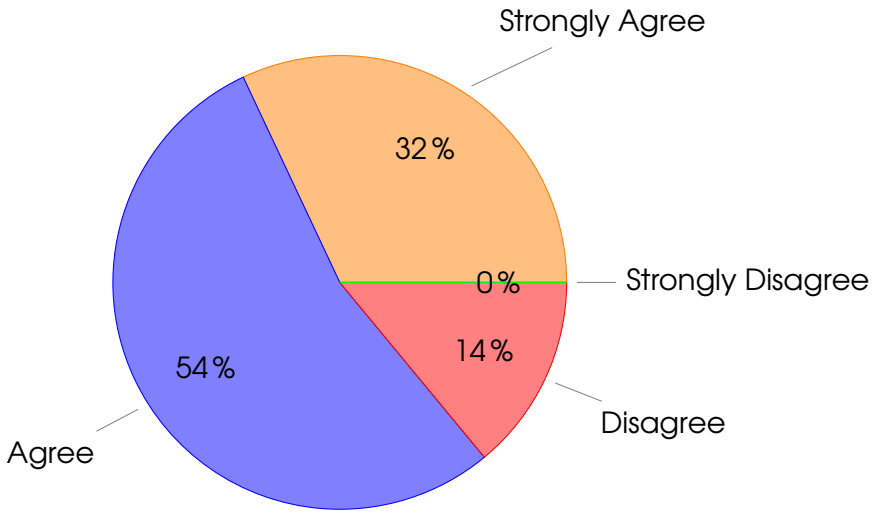


Figure 3.2: Results of the question: Do you agree with the following statement: ”I want to report suspicious e-mails sent to my professional e-mail account to an internal authority to to protect the organization from potential phishing attacks”

the correct authority. However, the value of any reporting tool strongly depends on the user satisfaction with the tool. This satisfaction largely depends on the usability and the feedback that can be acquired. Therefore, the participants were asked, how they would like to receive feedback to their suspicious e-mail reports. They had the options:

1. Weekly/Monthly organization wide information regarding the phishing schemes which are run against the organization
2. A personalized message informing me whether it was a phishing attempt or not
3. Only warnings from especially dangerous phishing attacks
4. No feedback at all

The responses to this question can be seen in Figure 3.3, where the numbers are the percentages of participants choosing this answer. It can be seen that most people want to receive either personalized feedback for each report or a month or weekly report about ongoing phishing schemes. Nevertheless, 22% of all participants only wanted to receive warnings about dangerous phishing e-mails. Another important point about the satisfaction with the feedback is the timeliness of it. Therefore, the participants were asked to specify the time span, in which feedback should reach them. The responses can be seen in Figure 3.4, which shows that about 7% of the participants would want to receive feedback within an hour. However, 43% would also be fine with answers within a day. But just as a many liked to only receive weekly or monthly reports. Last but not least, the participants which answered, that they would never want to receive feedback were also the same participants who wanted to receive no feedback at all.

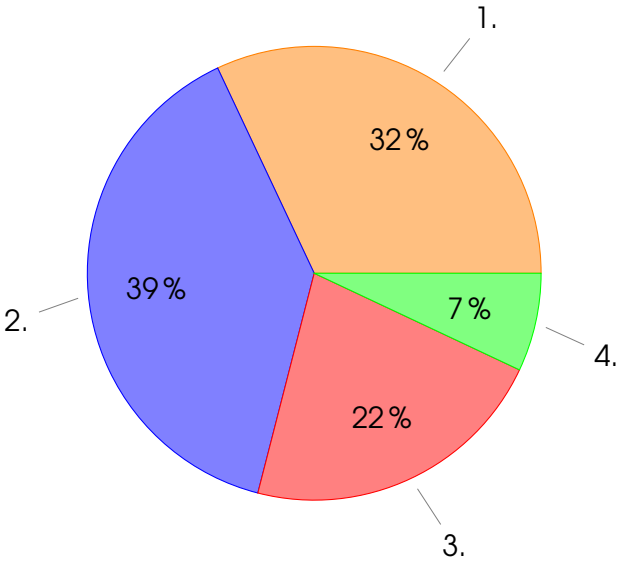


Figure 3.3: Answers to the question: How would you like to receive feedback

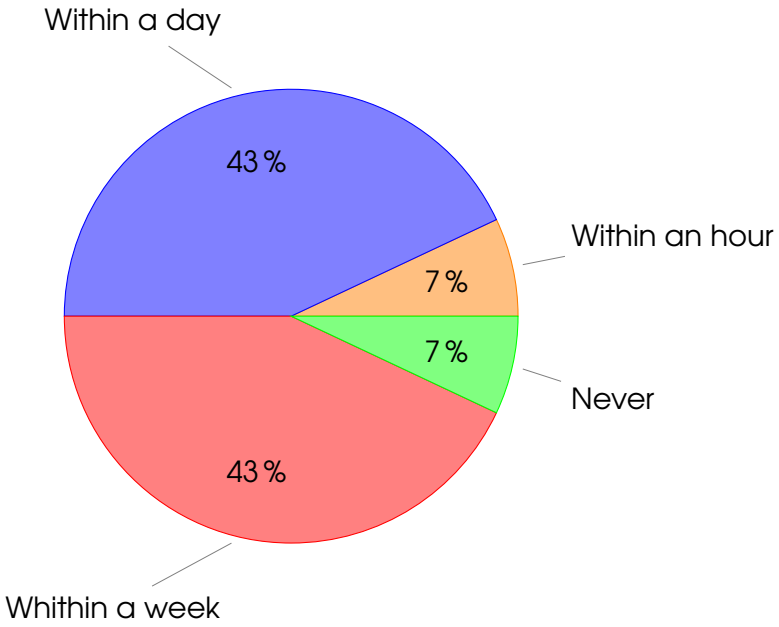


Figure 3.4: Answers to the question: How fast would you like to receive feedback

# 4 A Novel Approach in End User Reporting

Through the performed survey it can be seen that end users do want to report abusive e-mail, but due to a lack of knowledge or abilities, the majority of them can not report those e-mails. Therefore, in order to enable end users to report abusive e-mails, it has to be analyzed why they can not report these e-mails already. This is done in Section 4.1. Then it can be discussed how more end users can be enabled to report these e-mails, which is done in Section 4.2. Afterwards, in Section 4.3, a new reporting process can be defined, which uses the findings of the previous Section 4.2 to enable more end user to report e-mails. Furthermore, in Section 4.3.2 a plugin for the Thunderbird e-mail client is presented that implements the new process and thereby provides a proof of concept for its feasibility. Last but not least, Section 4.3.3 describes a tool for report receiver which can be used to process the reports generated through the new process.

## 4.1 Why End Users do not Report Abusive E-mails

During the report creation and sending, any user who wants to file an e-mail abuse report has to answer the following questions:

1. Should I report this e-mail?
2. What kind of e-mail abuse does this e-mail belong to?
3. To whom should I report this e-mail?
4. How should I report this e-mail?

At first sight, these questions seem trivial, however taking a deeper look at them reveals the difficulties end users can have with them.

The first question contains two parts. First a potential reporter has to identify an e-mail as being abusive. For some spam messages this can be quite simple, however spear phishing e-mails are a lot harder to detect, even for security experts. But it is even more difficult for less knowledgeable people to identify phishing mails. The reason for this is that they might lack the knowledge about e-mail headers, links and other technologies which might give a phishing mail away. This means that end users are quite often unsure whether some e-mail is a valid e-mail or a malicious e-mail. In this situation people might decide not to report such an e-mail due to the fear of falsely reporting an e-mail and thereby being punishable or potentially known as unknowledgeable. Thus potential phishing e-mails remain unreported.

Nevertheless, if the user has decided to report an e-mail, then the category of e-mail abuse which the e-mail belongs to, has to be specified. This is necessary because it has to be specified what kind of

incident occurred. In case of phishing e-mails, the incident category would be a social engineering attacks, while a spam e-mail belongs to the spam incident category. However, for end users the borders between the different incidents can be quite fuzzy. Furthermore, if the end user was unsure whether to report an e-mail in the first place he or she might give up, because the end user is not able to classify the e-mail with certainty.

For example, if an end user receives an unsolicited mail promoting certain online drug stores, this can be classified as spam. However, if the e-mail contained a link to a site which distributes malware, then it would be a phishing mail. Nevertheless, end users do not want to spend time investigating links in e-mails. Additionally, if they do not know how to safely investigate those links, they might even fall victim to the original attack by investigating the link. Therefore, end users are often unsure about the type of an abusive e-mail.

After the user has decided if an abusive e-mail is spam or phishing, the user then has to specify to whom t report this e-mail. Security experts can look up clearing houses and abuse contacts. The regular end user however does not know that such things exist, therefore the end users are often very limited in authorities to whom to report. Most often, they only know about authorities in companies they work for or other organizations they are a part of. Such organizations sometimes have security departments or CERTs, which makes it easier for members of these organizations to report e-mails, but even then it has to be decided whether the report should be send to the whole team or a specific address. Nevertheless clearing houses and security departments are not the only ones to whom reporting might be reasonable. Other reasonable authorities can be law enforcement, e-mail providers and other service providers, which are impersonated in the abusive e-mail, depending on the situation and e-mail. Through the huge amount of options and the limited knowledge of the end users, it is easy to see that end users willing to report an abusive e-mail have difficulties when choosing a report recipient, which again might discourage them from reporting those e-mails at all.

Last but not least, the reporting user has to decide how to report the e-mail. Many recipients accept reports of e-mails via e-mail, but there are often no instructions how such a report should look like. Security experts clearly understand that the report should contain the e-mail headers as well as the content and potential attachments, but the typical end user does not. Therefore, they often report just the e-mail content, which accomplishes little and thereby potentially frustrates them. Additionally this leads to a lot of information being lost in these reports, which reduces the reports value for the recipient. Moreover, even if the end user is told what the report should contain, many do not know where to find the necessary information and how to ensure that it is included. Thus making the report creation even more time consuming and complex for them.

Moreover, some of the information that could be included in reports might be sensitive for the reporter. In this case the end user would have to decide which information he or she can anonymize without harming the report. Furthermore, the end users would also need to understand how they can anonymize these information safely.

### 4.1.1 Organizational Context

Users who wish to report e-mails that are directed at their professional accounts face an even bigger problem. They do not only have to create a report that contains all the important information, but additionally have to decide which information the report is not allowed to contain. In the professional context many organizations have rules and guidelines declaring which information is sensitive and which is not.

For example, some organizations regard internal e-mail addresses as sensitive information. Thus in these organizations, the members would not be allowed to report e-mails from their professional accounts to outside authorities. Other organizations consider the IP addresses of their internal mail server sensitive, thus these IP addresses would need to be removed from the report.

Therefore, the end user would need to anonymize the reports and thereby remove the sensitive information before sending the reports. But this is very difficult, error prone and time consuming.

It can be seen, that the whole reporting process is quite difficult and time consuming. This poses another problem for regular users who want to report e-mail abuse, since the time they spend on reporting e-mails is time they cannot spend on their actual professional tasks. Therefore, they might decide to not invest in this difficult process and rather spend their time on other tasks. Hence, they simply delete the e-mails instead of reporting them, which results in the loss of information that could be gained from this e-mail and might have helped to defend from the attack or detect an already successful attack.

These difficulties are so big, that it is virtually impossible for the end user to create a meaningful e-mail abuse report in an acceptable time span.

## 4.2 How to Enable End Users to Report Abusive E-Mails

In this thesis, it is proposed that the end user should be enabled to send reports for **suspicious** e-mails, instead of helping the user identifying spam and phishing e-mails. The first advantage of this proposed solution is that the users do not have to identify abusive e-mails anymore, but rather can simply report any e-mail they feel unsure about. Additionally, the e-mails that could be identified reliably can be reported as well. The decision whether an e-mail is actually legitimate or not can then be made by an educated professional. In a professional and organizational context this professional can be an organization's internal authority, while external experts can be provided by clearing houses and other institutions for the private users as well as organizations. This also removes the second decision the end user has to make, which would be whether the e-mail that should be reported is a spam or phishing e-mail.

However, reporting suspicious e-mails forces the report recipient to investigate the reported e-mail more thoroughly to determine whether it is a spam, phishing or legitimate mail. But since report recipients have to validate any received report anyway, this additional burden should not pose a significant problem for clearing houses and other authorities. In organization internal authorities this might be different, if they did not process e-mail abuse reports before. But in this context it has to be assumed that the idea behind receiving and processing suspicious e-mail reports is to investigate spam and phishing campaigns against the organization. Thus these internal authorities can not prevent the added work

load anyway.
After a user identified an e-mail which he or she wants to report, the user has to decide on an authority to whom to report. In an organizational context this authority can be defined by the security guidelines and regulations, private users however have to look up potential authorities on the internet. In this thesis, it is proposed to simplify this step by providing the users with a list of authorities who would receive and process such reports. In organizations, this list would only contain authorities the organization approves of, such as their own CERT. On the other hand private users would receive a list, which would contain a variety of organizations that are willing to receive such reports from anybody.
The reason why such a list is important to enable end users to report abusive e-mails is that searching for an authority is time consuming. This will at least annoy end users, because reporting abusive e-mails is not their original focus. Therefore, they are unlikely to spend extended amounts of time with searching for suitable authorities. Furthermore, some end users might have problems finding such authorities at all, since they do not know what to look for. In suspicious e-mail reports this list becomes even more important as not every authority that excepts phishing or spam reports might accept those reports as well.
On the other hand providing them with a defined list of authorities limits the potential report receivers to the specified list. Thereby creating the need for cooperation between multiple authorities to spread the reported information further. Moreover, it has to be defined who can decide which authorities are on the list and which criteria an authority needs fulfill to be placed on the list. Last but not least, if an authority decides to stop accept those reports, or a new authority would like to receive such reports, the provided list needs to be updated regularly. This raises the question how to do this securely.
The best place where such a list could be located is the e-mail client of the end users. The reason for this is that the end users will already be using the e-mail client when they read their e-mails and thereby potentially identify suspicious e-mails. This provides them with the necessary information when the end user needs it.
The last step before sending the report is to create the report message. As described previously, the difficulty lies in collecting all information necessary for processing the e-mail reports. Since this information is clearly defined as all the information in the e-mail headers as well as the content, these information can be obtained automatically. However, some information among the headers and inside of the e-mail might be confidential or personal. Therefore, the reports should not be automatically sent, rather the end user needs to be able to read what he or she will be sending and potentially anonymize it. The anonymization is quite difficult to perform automatically for all users, since different users define different information as confidential. Thus the end user has to perform anonymization manually. In organizations however, the organization can create an internal authority, which then can handle the anonymization process for its members. This alleviates the end user form the burden of removing all sensitive information in the report and removing the risk of data leakage through anonymization errors. Internal e-mail abuse reports have further advantages for organizations. First and foremost, internal reports notify an organization directly when it is targeted by a phishing attack. Additionally, this allows them to evaluate reports, which can generate information about the volume of phishing attacks in

general as well as information about potential goals of attacks. Furthermore, they enable the organization to internally investigate if the organization has already fallen victim to the attack and to prepare defenses against the attack if not.
The disadvantage for an organizations is however, that they potentially have to deal with a huge amount of abuse reports. This can become very expensive if each message has to be evaluated manually. However, if the reports are all formatted in a machine readable format, the amount of effort required to handle the reports can be reduced drastically.
The first step in managing the amount of reports is to automatically detect duplicate reports. Such duplicate reports are multiple reports send for e-mails which contain the same content. This can occur if an end user reports the same e-mail twice or if multiple users received similar e-mails. If such a duplication is found, the investigators just have to look at one e-mail, identifying it as legitimate or not. These duplicate reports should however not be deleted as they might give information about the targeted individuals of an attack, the goal of the attack and the attack volume.
Furthermore, links in the reports can be automatically checked against publicly available blacklists of phishing sites, which might classify a certain amount of reports completely automatically.
Moreover, public authorities would also benefit from the automation potential of a standardized machine readable reporting format. However, if the end users have to create the report manually, such a standardized reporting format can hardly be ensured, as every end user would have to know the format and be able to create reports in it reliably. But if the reports are created automatically, as suggested above, using a machine readable format becomes a lot easier. Nevertheless, it has to be prevented that a recipient has to be able to parse the exact format as this would limit the range of recipients. This can be ensured by choosing a format, that is human readable as well.

## 4.3 The new Reporting Process

Using all these thoughts on how to enable end users to report abusive e-mails, allows for the specification of a reporting process that is user friendly but at the same time creates highly valuable reports for the receiving authorities. This reporting process can be seen in Figure 4.1. In the Process seen, diamond nodes are decisions the end user has to make. Squares represent actions the user has to perform and the red ellipses are things that happen automatically. The reporting process will now be walked through using a generic reporting assistance tool for the automated parts. In Section 4.3.2, an Add on for the Thunderbird e-mail client is presented, which was developed during this thesis as a proof of concept for the feasibility of such automated tools.
At the beginning of the process is the reception of an e-mail. When the end user looks at the e-mail, he or she has to decide whether the e-mail seems suspicious. An e-mail can be suspicious because it was received unsolicited or because of its content or sender appearing suspicious. Additionally, e-mails with offensive content can also be counted as suspicious. The user should then tell the automatic assistance tool that the e-mail is suspicious.
This will then lead the reporting tool to automatically parse the information required for the report from the e-mail. The reason why the information gathering phase is completely automated is that end

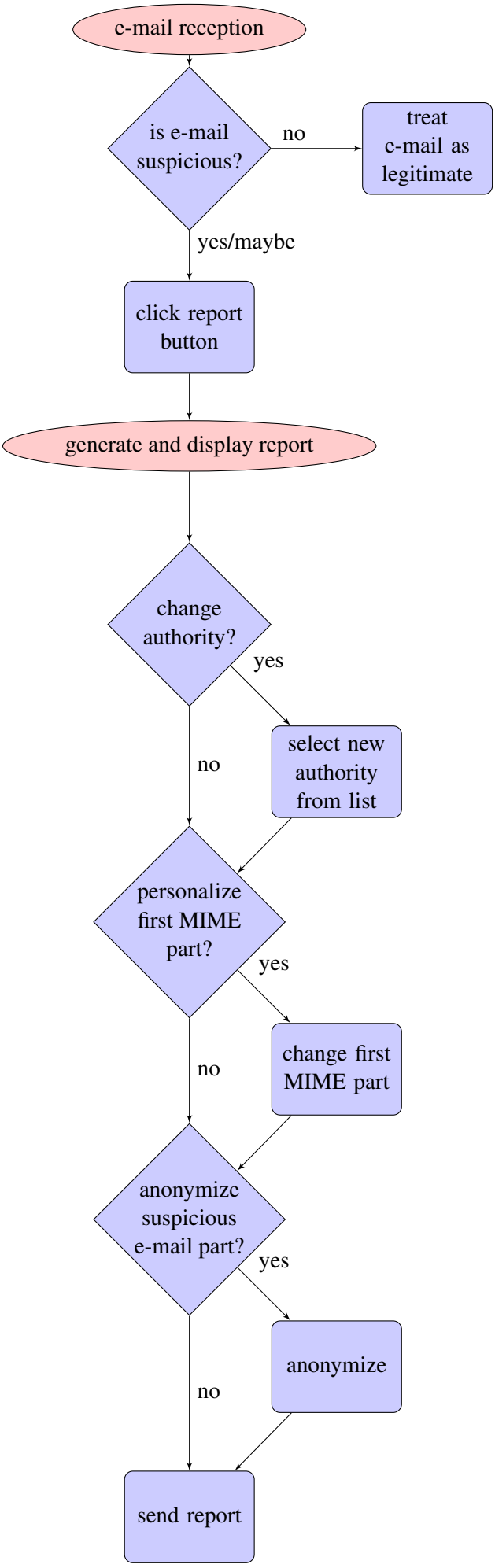


Figure 4.1: E-mail Abuse Reporting Process with Plugin Support

user generally do not know which information to include in reports and where to find them. Furthermore, even if the end user knows which information to include, it remains time rather time consuming to extract the information form the e-mail manually. Therefore, automation makes the report creation easier as well as faster and at the same time enables non tech savvy end users to create better reports.
On the other hand, completely automating the report generation prevents the end users from being able to decide which information they want to share with the authorities. Thereby raising confidentiality and privacy concerns. These concerns stem from the fact that legitimate e-mails as well as spear phishing e-mails often contain private information, such as names or e-mail addresses, which could be part of the automatically parsed information. Additionally, the end user could have identified certain suspicious features in the e-mail which he or she would want to point out to the report recipient, which is not possible in completely automated reports since the user has no control over the included information. Therefore, it is important to give the end user the possibility to anonymize the report before sending it and allow him or her to add information to the report if necessary.
Since the report is automatically generated, it is sensible to choose a standardized report format for the reports. In Section 4.2 it was already discussed, that such reports should be formatted in a format which is both machine and human readable format. The creation of a new format for this purpose is not sensible for two reasons. The first reason is that creating a new format for each new purpose forces recipients to deal with a huge variety of formats for the different reports. This reduces the usefulness of automatic analysis tools as well as puts a huge burden on organizations which want to process reports. Additionally a huge variety of reporting formats discourages standardization and complicates cooperation between different organizations.
Secondly, there are already formats that are human and machine readable, which are also extremely versatile. Therefore, a new reporting format would be obsolete. In a short analysis of the different available formats, see Section 2.3, X-ARF seemed to be the perfect fit for suspicious e-mail reports in the new reporting process. The first major advantage X-ARF has over other formats is its variety in incidents that can be reported using this format. This allows for an easier integration at the receiver side of the report. Additionally, since X-ARF reports are based around e-mail communication, they can be sent via e-mail from the mail client. Furthermore, displaying the report can be performed in the same way as standard e-mails are created in the specific client. The anonymization could then be performed in a similar fashion as the creation of regular e-mails, which is familiar for the end users and thereby easier for them to use. Last but not least, X-ARF allows for a free specification of the reported incidents. Therefore, it is possible to specify a report type for suspicious e-mail reports, which is specifically designed to contain the information that is useful for classifying e-mails. This also allows organizations to specify their own reporting format for internal mail reports, which gives them the option to include more information and organization specific information in their reports. If the reports are sent to external authorities however, it is sensible to create a single standard schemata that everybody uses. This allows for multiple institutions to be notified with the same report as well as the exchange of report information between different organizations. This external standard format has to include all necessary information for report processing, however it should not include any information

that could be considered confidential. This makes the report format for reports to external authorities usable for internal reports as well. Furthermore, the external schemata can be used as a blue print for a potential internal report by simply adding the information not covered by the external report.
As the Section 2.3.1 describes, X-ARF consists of three MIME parts. In the suspicious e-mail reports, these MIME parts are used according to the X-ARF standard in the following way.
The first MIME part is filled in with an automatically generated short human readable text, which simply states what is reported and what report format is used. The information contained here are the date of the report and the e-mail server the e-mail was received from. This part is also perfect for the end user to provide comments to the report and enables the end user to add information to the report, just as requested previously. For example users can state why they find this e-mail suspicious in this part.
The second MIME part contains further information about the report. The information in this part are specified in the used report schema. In Section 4.3.1 a report schema for external reports is proposed. All the information in this MIME part are automatically read from the reported e-mail and formatted according to the YAML specification. Apart from optional fields in the schema, the end user should have no influence on the information in the second MIME part.
This is important since end users can not be expected to understand the specific formatting requirements of the machine readable part. Therefore, end users are likely to invalidate the format. Furthermore, end users do not know which information is vital to such reports and could thereby remove information that is explicitly required by X-ARF reports, which would break the report entirely.
At the same time this raises privacy concerns regarding the information in the non optional fields, as these information can neither be anonymized nor excluded. To resolve this it is important to design the used schema to only state non sensitive information as required.
The third MIME part in suspicious e-mail reports contains the e-mail that is reported. The reason why the whole message is appended to the report is that firstly any authority, especially external ones, need some kind of proof that the reported incident has really taken place before they can act. Therefore, appending the reported e-mail is a very good proof for the reception of a suspicious e-mail. Furthermore, the information in the second MIME part are only the most basic information for determining the validity of an e-mail, especially if all optional field are omitted. Therefore, the reported e-mail is required by experts as a source of information for determining the legitimacy of an e-mail.
However, adding the reported e-mail to the report might raise some privacy concerns, as the e-mail might contain personal data. This has to be taken seriously as among the reported suspicious e-mails, will most likely be some legitimate e-mails. These are especially likely to contain sensitive information about the reporter or the innocent e-mail sender. Therefore, the reported e-mail has to be anonymizable. But the amount of anonymization should be kept to the bare minimum since to much anonymization might make the report worthless to the recipient. On the other hand anonymizing organizations secrets and personal information sometimes might allow the report to be shared with a wider range of authorities, which multiplies the effect of the report. Therefore, it is recommended, to anonymize as little as possible, but as much as necessary.
The e-mail should also be automatically copied into the third MIME part. This has the advantage that

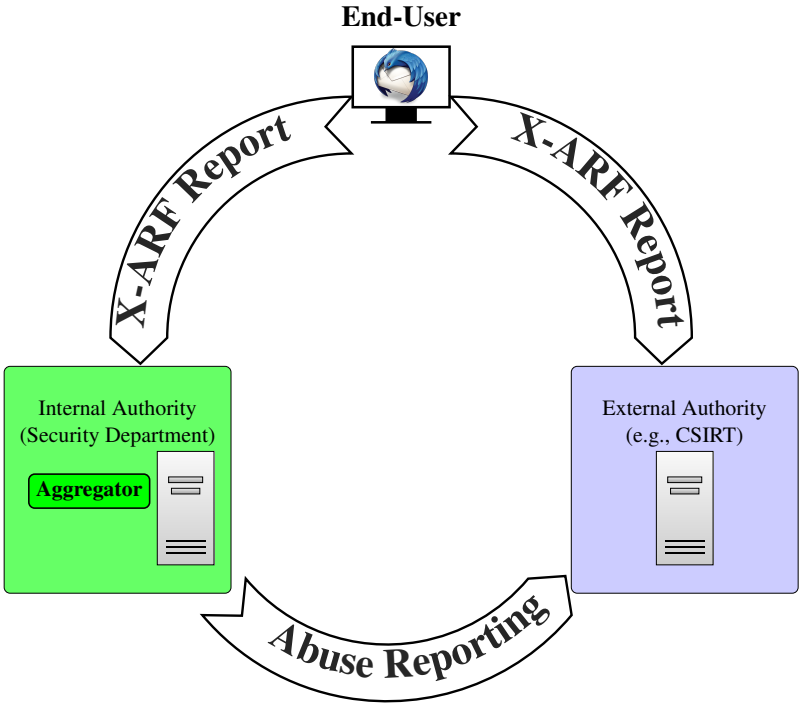


Figure 4.2: Reporting Procedure Overview

the whole e-mail, including all headers and attachments, are appended and not only the e-mail content, which often happened when the end user did this manually.

After the report was generated automatically it should be displayed to the end user, such that he or she can read what is reported before sending it. At this moment, the authorities address to whom the report should be sent to should already be selected and presented to the user in an address bar.

The next step for the user is now to look at the selected authority and verify that the selection is correct. This should usually be the case, however if not the user is supposed to select another authority from the provided list of authorities. The next step in the reporting process is then for the user to decide whether he or she wants to personalize the human readable part of the X-ARF report. If this is the case, the user modifies the content of the first MIME part. If the first MIME part should not be personalized, the user can directly move on to the next step.

This next step is for the end user to decide if he or she has to anonymize the reported e-mail in the third MIME part. Again, if anonymization is required, the user can anonymize the sensitive data by editing the third MIME part. If no anonymization needs to be performed, the user can directly move to the last step.

The last step for the user is to send the report. At that point the report should be sent to the selected authority. This authority can be internal or external, as seen in Figure 4.2. After sending the report, the reporting tool can perform certain actions that provide instantaneous feedback to the user. Among those actions is marking reported e-mails, which gives the user feedback on what he or she has done and at the same time prevent duplicate reports. An example implementation for such a reporting tool can was developed for the Thunderbird mail client and is described in Section 4.3.2.

On the receiver side, the report can then be read and verified automatically. Bigger public organizations such as phishtank and CERTs usually have their own automated report validation and processing. Organizations, which prior to the new reporting process did not engage in e-mail abuse reporting, most likely will not have such an automated process. However, the X-ARF format of the report enables a very simple implementation of automation for these reports. As a proof of concept, a report aggregator was developed. The report handling process on the receiver side with such an aggregator is described in Section 4.3.3.

### 4.3.1 The Report Schema

The report schema in X-ARF is extremely important as it defines all information included in the report. Therefore, different organizations might want to include different information in this schema. This is no problem as long as the reports are send organization internally. However, reports to external authorities have much stronger restrictions regarding potentially sensitive information. Therefore, it is important for a schema to allow for the exclusion of these sensitive information. In this thesis a report schema for external reports is proposed. The proposed schemata can be found on the CD part of the appendix of this thesis.
The external schemata inherits the required fields described in Section 2.3.1. But poses limitations on the potential values for some fields. This is done to fit the contained information to suspicious e-mail reports. The specific limitations and fields will now be discussed individually.

**Schema-URL:** The Schema URL field contains the URL of the schemata for the abusive e-mail reports external schemata. Therefore, it should be an URL that points to a publicly available schema definition, in order to allow everyone to read the specification and validate the report. The existence of this field is required by the X-ARF standard. The values in this field are not restricted, to allow the schema location to be moved more freely. This is also has the advantage, that multiple organizations can host the schema. Thus if the main host of the schema moves or is taken offline, organizations which host the schema as well can still use reporting tools based on the schema without having to modify the tool.

**Version:** The version field specifies the X-ARF version that is used in the report. This is necessary to allow for multiple generations of the X-ARF standard to be separable from one another. This version highly depends on the reporting tool that generates the report. The report schema is designed to be used in conjunction with the 0.2 X-ARF version. However, to support potential new versions as well this field remains unrestricted.

**Reported-From:** The X-ARF standard requires this field. It should always contain the reports originating address. Therefore, the reporters e-mail address should be found here. The potential values in this field are only restricted by the e-mail address format. This is required by the X-ARF specification.

**Report-ID:** Again, in every X-ARF report, the *Report-ID* field has to be present. It contains a unique identifier for every report. This identifier can be created by combining the current time stamp

with the reporter's e-mail address, hashing it and then appending the e-mail domain of the report sender to the result. This will result in an id of the following form <*random number*>@*domain*, which should be resilient enough against collisions and unique over domain and time.

**Report-Type:** This field is also required by the X-ARF standard. However, the external suspicious e-mail report schemata requires it to contain the key *suspicious-e-mail*. This enables automatic tools to categorize the report correctly. Furthermore, it provides recipients with a unique name for the received report, which they can look up to correctly interpret the information in the report. Since this string is supposed to be a unique name for the report, this field has to be restricted to a single enum string.

**Occurrences:** The occurrences field in the X-ARF specification is used to report multiple similar incidents at once. However, as stated in Section 2.3.1, each incident should be reported individually to ensure that deviations in the incidents as well as the information regarding the timing of the incidents remains intact. In suspicious e-mail reporting, this field should only be used when reporting the reception of multiple identical e-mails. But this hardly ever happens because in most cases, the intended reception e-mail addresses differ. Therefore, sending a report for each variation of the *To* field is advised as only then can the attack target be judged accurately by professionals. Nevertheless, it might be handy to use the occurrences field in order to preserve the scale of an attack, while masking the individual targets, by reporting a single mail and setting the occurrences field to the amount of messages that were received in total.

**Category:** This field is also defined in the X-ARF standard. But for suspicious e-mail reports, the potential field values, described in Section 2.3.1, can be drastically reduced. First of all suspicious e-mail reports will never be used to indicate an *auth* incident as suspicious e-mails rarely pose an indication for authentication failure or abuse. Nevertheless, suspicious e-mails could fall both under fraud and abuse. However, both categories suggest that the reporter is sure, that an actual attack or incident has taken place. But suspicious e-mail reports can be reports about legitimate e-mails, thereby no abuse or fraud has taken place. This leaves only the two categories *info* and *private*. The *private* option indicates that the report should only be viewed by the reporting and receiving party and nobody else. This limitation might make sense in the context of internal suspicious e-mail reports and therefore might be a good modification for internal schemata, however in external reports, the report efficiency profits, if the report can be shared. Therefore, the external suspicious report schemata restricts the report category to *info*. This is fitting as each suspicious e-mail report is basically an informational report that abusive e-mails might have been received. Whether such a report describes a real incident can only be unambiguously said after evaluating the report. Furthermore, if the suspicious e-mail report contains information that the recipient is not allowed to share with others, then the *TLP* field can be used to restrict the information propagation.

**User-Agent:** This field is supposed to contain information about the tool used to create the report.

The string that this field contains is solely dependent on the reporting tool that is used. Therefore, it makes no sense to try to restrict this field. However, when designing a reporting tool, a user agent string should be chosen which is rather unique, in order to clearly identify the tool used to create the report. Otherwise, it is impossible for the recipient to tailor a response to the reporting system. Furthermore, it becomes impossible to identify broken implementations which create invalid reports.

**Date:** X-ARF specifies that the *Date* field should contain the date on which the incident was noticed. In suspicious email reports this would be the date at which the e-mail was received, read and found to be suspicious. Therefore, the reporting tool should automatically fill in the date and time when the report was generated. As the format for the date, the X-ARF standard recommends the RFC 3339 format. However, due to compatibility reasons the standard also accepts the old RFC 2822 date format as well. Therefore, all report receivers have to be prepared to receive dates in the old date format as well. The schema could try to enforce the new time format, since any suspicious e-mail reporting tool will be created after the X-ARF 0.2 standard was published. However, enforcing this time format would only make the development of new reporting tools generating suspicious e-mail reports more difficult, while at the same time all recipients of X-ARF reports still have to be able to handle the old date format.

**Reception-Date:** This field is unique to the suspicious e-mail report schemata. It contains the date at which the reported e-mail was received at the last hop. This separate time field is necessary as the incident detection time and the e-mail reception time can be days apart. However, including the reception time allows external authorities to notify mail servers that were abused to send abusive mails about the time at which the mails are send. This can act as proof for the reported incidents, because the mail provider can study the logs and thereby reconstruct the attack. Furthermore, authorities can correlate the reception date with the reception dates of other similar reported e-mails. This allows them to estimate the size of spam and phishing attacks more precisely. Internal authorities can even use the reception date to start forensic analysis in order to determine if the attack was already successful. The format in which the date is provided is the same as in the previous field and regarding the restriction of the date format to a single standard the same argument is valid as well. Therefore, dates in RFC 3339 and RFC 2822 formats are accepted.

**Source:** The source field in X-ARF is supposed to contain information about the source of the incident. In suspicious e-mail reports this is rather tricky, because sender information as well as mail server hops can be spoofed. However, in the external schema, both the e-mail address from the *From* e-mail header as well as the first mail server hop are accepted sources.

In other X-ARF reports, this field is extremely important and thereby very restrictively defined regarding the content. However, spam as well as phishing e-mail do often conceal the original source of the message. Therefore, it is quite challenging to define a good source for those e-mails. If one uses the e-mail address in the *From* field this directs the attribution towards an e-mail address. This can be quite useful when dealing with spam messages, because spammers

do use automatically generated or compromised accounts, the used e-mail address is often the actual source of the e-mail.

In contrast, phishing mail sender often forge the *From* field. However, they sometimes do not fake the e-mail server hop history. In those cases, the originating e-mail server IP can be seen as the source of the attack. If both the *From* field as well as the mail server hops are forged, then experts are needed to evaluate the e-mail headers in order to find the origin of the message. This is very complex to do automatically. Thus in this schema it is recommended to enter either the *From* or the e-mail server's IP from which the e-mail was received as source. But the value in this filed is in no way restricted to only those two options. Thus if a reporting tool exists that can reliably determine the true origin of an e-mail, it can always provide the correct incident source in the source field. However, recipients have to treat this value with caution.

**Source-Type:** This field is used in X-ARF reports to determine what the *Source* field contains. Therefore, a fixed set of specified source types is defined. However, for suspicious e-mail reports only IP addresses and e-mail addresses make sense at all. The other two options *uri* and *domain* are either not fitting the potential sources or are too broad. The *uri* specifier is supposed to be used when a link to a specific resource is the incident source, this is applicable for example in malware reports but not in suspicious e-mail reports. The *domain* specifier would on the other hand mean that a complete domain such as *@gmail.com* would be given as the source. This would be far to broad to give any indication about the real sender of an e-mail.

**Mail-Server-Hops:** Another field which is unique to suspicious e-mail reports is the *Mail-Server-Hops* field. It is an optional field, which contains a list of the mail servers the reported e-mail supposedly passed through. This list is not always the actual path of the e-mail as it is possible to spoof the involvement of some mail servers. The list is ordered in the same way as the *Received* e-mail headers are which means that the first mail server is at the end of the list, while the targets mail server is at the start. In some occasions, the IP addresses of internal servers are deemed confidential. Therefore this fields is chosen to be optional. Thus, this field can be omitted from a report and the sensitive information can be anonymized without manipulating the data in the second MIME part. The reason why this is anonymization inside of the JSON object in the second MIME part is not advised, is that the report format could be broken and therefore, the report made useless. Furthermore, end users are supposed to report the suspicious e-mails and most of them do not know which of the IP addresses belong to internal servers. Therefore mistakes in the anonymization are very likely which makes excluding the received addresses the better option. Should the mail server hops be excluded in the second MIME part, the reporter can still decide to anonymize them in the third MIME part or to exclude them completely. This will however most often lead to the report being discarded and is not advised.

**URLs-Found:** This is another optional suspicious e-mail report exclusive field. It is supposed to contain a list of all unique links in the suspicious e-mail. The list of links makes the identification of phishing e-mails a lot faster. The reason for this is that these links can be checked against black

listed sites, which automatically can classify some e-mails as phishing. Furthermore, experts can very often judge the validity of an e-mail by the links alone, because links containing misspelled URLs or IP addresses are most often phishing mails. However, sometimes in targeted phishing attacks, the mail might contain sensitive links. Such sensitive links can for example be links to internal services that an attacker might have found but an external authority should not gain knowledge about. Furthermore, if the suspicious e-mail was not in any form an abusive e-mail, the links included in the e-mail might contain confidential information for the organization reporting that e-mail. Thus they can exclude the link information in the second MIME part and potentially anonymize the links in the third MIME part.

**E-Mail-Addresses-Found:** In this optional suspicious e-mail exclusive field, all unique e-mail addresses that are found in the reported e-mail are listed. This list of e-mail addresses can be used to automatically judge reported e-mails. For example, if accidentally or purposely organization internal legitimate e-mails are reported, they should only contain internal e-mail address. This makes reported e-mails with solely organization internal e-mail addresses likely to be legitimate. At the same time, it is possible to use blacklists with e-mail addresses, or whole e-mail domains, to automatically classify some e-mails as phishing or spam e-mails.
Furthermore, such a list can be used by experts to classify e-mails as well. This can be done for example by searching for deceptive e-mail addresses, such as slightly misspelled ones.
Nevertheless, in some cases e-mail addresses contained in a reported e-mail can be sensitive. Therefore, it is has to be possible to exclude the e-mail addresses in the second MIME part, which allows the reporter to anonymize the sensitive data in the third MIME part before reporting the e-mail.

**Feedback-Address:** The last unique field of the suspicious e-mail report schemata is the *Feedback-Address* field, which contains an optional feedback address. This address can be equal to the report sender address but does not have to be. The survey, that was performed in this thesis (see Section 3.7) revealed, that some people (7% of all survey participants) do not want to receive any feedback, while others only want to receive feedback once a week. These people can leave the feedback address field empty or insert an address which gathers the feedback and creates warnings and reports for those users. Alternatively the field can also be omitted which means that the report sender does not want to receive feedback as well.

**Attachment:** This field specifies which type of content is presented in the third MIME part. For suspicious e-mail reports, this will always be the complete reported e-mail. Therefore, the only valid value for this field is *message/rfc822*. The schema requires that the e-mail is attached as only this e-mail can provide enough evidence to prove that an incident has taken place as well as prove that the claimed incident source is the actual incident source. Furthermore, in this e-mail are also information, which might be required to classify the e-mail with certainty. Both cases become even more important if the optional fields in this schema are omitted. In this case nearly no information about the reported e-mails are given, which makes it impossible to classify the

e-mail correctly.

**TLP:** Last but not least, the *TLP* defines the sensitivity of the information in the report. It should be chosen to be as restrictive as necessary, but also as liberal as possible, to enable collaboration between the different authorities. This can maximize the impact of each report.

The schema is proposed as a schema for reports to external authorities. However, since most information in the schema are also of interest to organization internal authorities, it can also be used for internal reports. But reports to internal authorities do rarely have to adhere to the same confidentiality and secrecy constraints. Therefore, these reports can include sensitive data as well as information which are only relevant in this specific organizational context, such as department codes. In this case, the internal authority can define their own schema, which can be based on the proposed one, for their individual suspicious e-mail reports.

### 4.3.2 The Thunderbird Suspicious E-mail Report Add-on

In this thesis a Thunderbird Add-on was developed, which can serve as an example for the reporting tool used in the Section 4.3. Thunderbird Add-ons are developed in JavaScript and can inject UI elements into the Thunderbird interface through overlays.

The first question in the development of any reporting Add-on is how the users should interact with the Add-on. Answering this question is quite easy for such a reporting tool, since the easiest form for any user to interact with any system is a simple button press. Therefore, the Thunderbird reporting plugin defines a button inside of the Thunderbird UI, which starts the automatic report generation. However, this raises the question where to place this button. In Thunderbird, there are many possible locations to locate the button, for example a variety of tool bars and button bars. But the most sensible place to locate the report button is the button bar which contains the other e-mail interaction specific buttons, such as *Delete* and *Reply*. This can be seen in Figure 4.3.

The reason why this specific location was chosen is that users of the Thunderbird mail client are used to finding all the functionality related to a single e-mail in this toolbar. Thus the usage of the plugin is more intuitive, if the button is placed there. Furthermore, any Thunderbird user knows that buttons in this button bar perform an action regarding the currently selected e-mail. Therefore, users will intuitively know which e-mail they are currently reporting.

Additionally, if an end user finds an e-mail suspicious, he or she do not have to search around for the report button, rather they can immediately spot it at an expected location. Finding the button is made even easier through the usage of color in the logo on the button, while the other buttons are kept in black and white. Obviously these specifics are only valid in the Thunderbird mail client, but the principle that the report button should be located alongside the other e-mail related buttons as well as the improved visibility increase the usability of any e-mail reporting tool implementation.

Hence, if an end user perceives an e-mail as suspicious, he or she simply clicks on a report button. Once the *Report* button is clicked, the plugin will load the e-mail headers out of the Thunderbird e-mail database. Using these headers the Add-on is able to locate the complete e-mail and load it asyn-

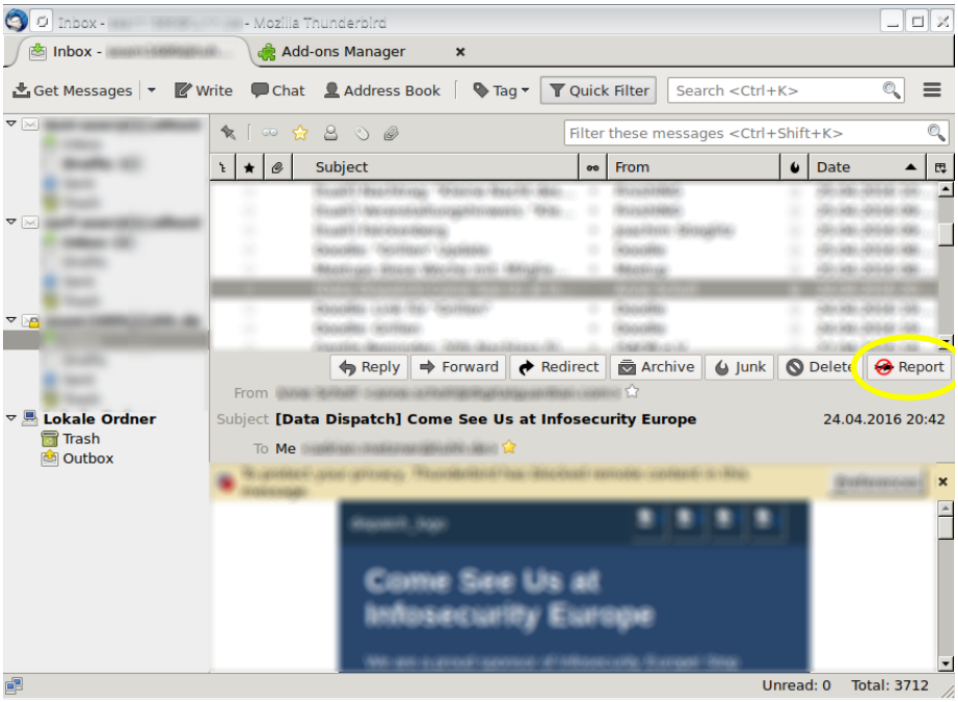


Figure 4.3: Thunderbird E-mail Client with the Report Button

chronously. The reason why it has to be loaded asynchronously is that e-mails can become extremely large. To enable Thunderbird to run smoothly and not behave unresponsive over longer periods of time, Thunderbird forces any access to the actual mail to be performed asynchronously.

When the e-mail was loaded the plugin starts to analyze it. First and foremost, it calculates an id for the report. The id is calculated by obtaining the current time stamp, appending the users e-mail address and hashing it. Afterwards the users e-mail domain is appended to the result. This creates a unique identifier for the reports which follows the creation proposal described in the *Report-Id* field description in Section 4.3.1.

Then the Add-on starts reading and interpreting the e-mail headers. In those headers the Add-on can find the *From*, *To*, *Date* and *Received* headers. The *From* and *To* fields contain potential e-mail addresses that can be useful for the *source* field in the reporting schema as well as the list of e-mail addresses contained in the e-mail. However, in the current version of the plugin, they are solely used in the list of found e-mail addresses. The reason for this is that another value is considered more reliable as an incident source.

The *Date* header contains the reception date of the e-mail. This is important for the *Reception-Date* field in the report schema. The *Received* header contains a string in which all mail servers IP addresses, which the e-mail traveled though, are noted. Since this string contains many more information apart from the IP addresses, the Add-on takes the string and searches for the servers IP addresses and places them in a list. This list is ordered in the same way that the received header is ordered, which means the last entry of the list is the mail server IP from which the e-mail originally came. The reason why this ordering is kept is that this ordering has become a standard for experts. Therefore, changing this

standard might lead to confusion on the report receiver side.
Furthermore, the Add-on searches through the complete e-mail looking for links and e-mail addresses. If it finds one of these, they are read out and placed in a list with other e-mail addresses or links respectively. After searching through the entire e-mail, the Add-on has created two lists, one full of e-mail addresses contained in the e-mail and the other full of links contained in the e-mail as well. Then the Add-on goes through the lists again to purge them from any duplicates.
After gathering all the data, the Add-on automatically creates a report, just as the new suspicious e-mail reporting process demanded. Therefore, the Add-on creates a new e-mail with the required X-ARF headers, meaning X-XARF: PLAIN and the correct content type. The *From* field is filled with the users configured identity which is store in Thunderbird. Furthermore, the subject of the e-mail is set to *Suspicious E-mail report <id>*.
Then the first X-ARF MIME part is assembled by creating a text based e-mail MIME part using a standardized human readable message, in which it inserts the reported e-mails source as well as the reception date.
Afterwards the plugin creates the second MIME part by filling in the required information in the fields defined by the report schema. Most of the fields are straight forward, however in the source field the plugin currently uses the first IP in the mail server hop list, which is the last mail server before the e-mail reached the destination e-mail server. The reason why a mail server IP address was chosen is that such addresses are a lot harder to spoof. The last mail server before the destination server was reached was chosen as a source, because it is the only IP address in the whole mail server hop list that can not be spoofed. The reason why it can not be spoofed is that it is written into the header by the destination mail server. Therefore, to spoof this IP address, the attacker has to gain control over the e-mail server which the reporting user uses. This is extremely unlikely. The disadvantage of this decision is that the report will rarely contain the original source of the suspicious e-mail but rather only a server redirecting the mail. However, certain authorities may be able to follow the e-mails path through the mail severs backwards to the non spoofed originating mail server.
Since the e-mail uses an IP address as a source specifier, the source type field has to be set to *ip-address*. Furthermore, the dates in fields *Date* and *Reception-Date* are both formatted to adhere to RFC 3339, in order to comply with the newer recommendation of the X-ARF standard and the user agent string field is set to *x-arf-reporter-plugin-thunderbird/0.1*. Last but not least it has to be noted, that the filed values for the *TLP* and *Feedback-Address* fields are read from the plugin configuration. This configuration can be changed by the user any time he or she likes.
At last, the Add-on fills the third MIME part automatically with the reported e-mail.
After the report was automatically created, it is displayed to the user in a window, that almost resembles a normal e-mail compose window, which can be seen in Figure 4.4. This ensures, that the end users know how to interact with the window without being taught. The window is defined in the plugin as a completely separate window, with just the look and feel taken from the regular compose window. This allows the plugin to be more freely able to adjust certain aspects in the window. The most important aspect which can be achieved through the usage of a new window is the separation of the compose

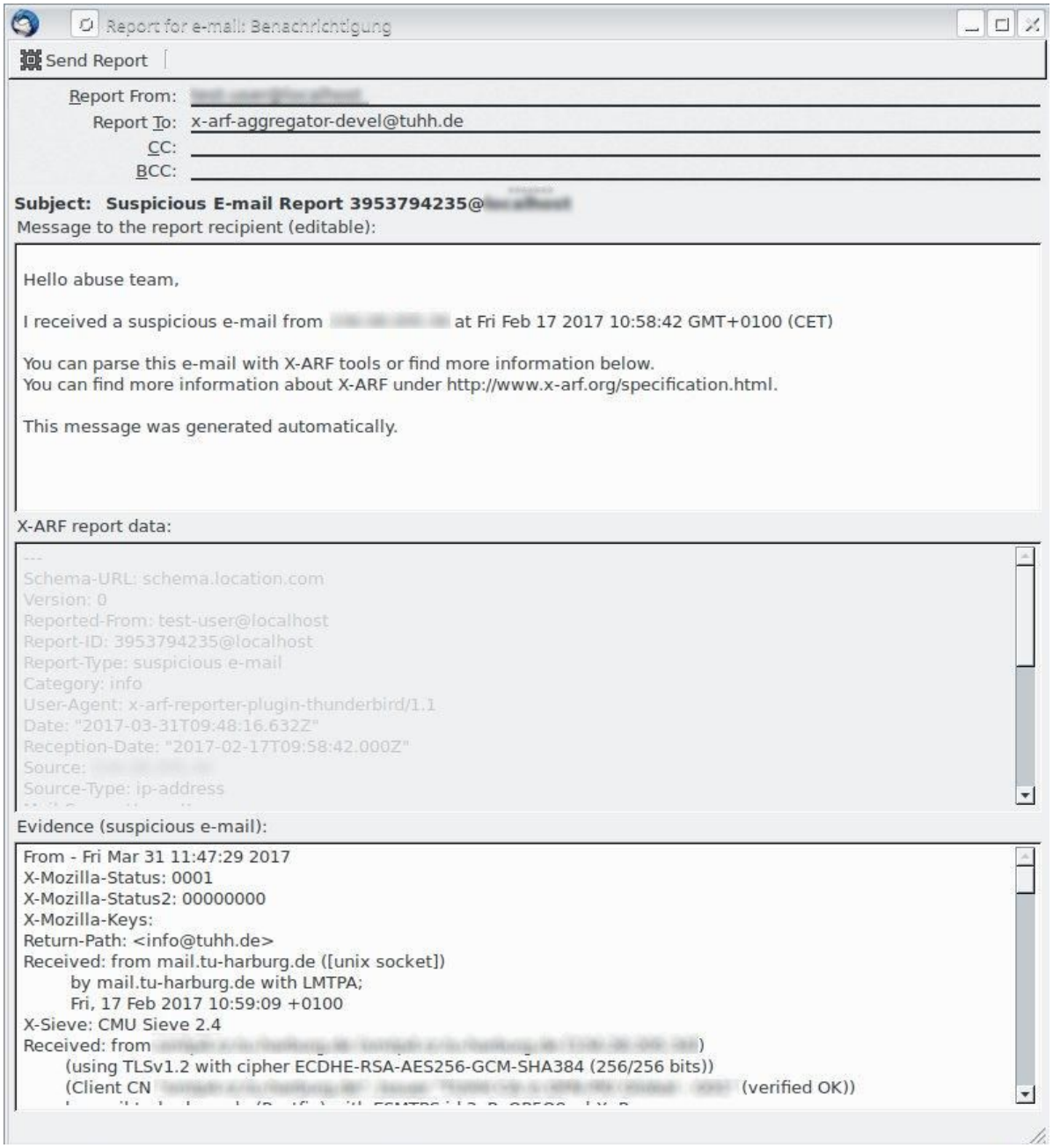


Figure 4.4: Thunderbird E-mail Abuse Report Compose Window

area. The compose are is in the report window separated into three parts. Each of these parts contains exactly one of the three MIME parts of the X-ARF report. There are two reasons why this split was made. The first one is simply that splitting the compose window shows the user that he or she will send a report which consists of three parts. Thus without knowing about the specifics of X-ARF the user understands that he or she will be sending these three parts in the report. Additionally, this also allows the end user to read and manipulate the report parts he or she is able to manipulate separately from one another. Thereby the user is unable to break certain necessities of the MIME standard by accident.
The second mayor advantage of splitting the compose area is that each part can be made editable or unmodifiable on its own. This is used to disable user modification of the second MIME part in the X-ARF standard by simply making it unmodifiable. The reason why this part should be unmodifiable is given in Section 4.3. However, the first and third MIME part still remain editable. Therefore, the user can still add information to the first MIME part and anonymize sensitive parts of the reported e-mail.
At this time the e-mail address of the receiving authority is already filled in as the e-mail receiver. Which authority is chosen is defined by the user. The user can choose the authority in the options of the Add-on. If no authority has been explicitly chosen, the first authority in a predefined list of authorities is chosen. This list is also the list of available authorities the user can choose form. It is supplied as a hard coded list in the Add-on. The original design of the Add-on would have used a list provided in a separate file to load the list of authorities on the start-up of Thunderbird. This would have the advantage that it becomes very easy to update this list as well as allow organizations to easily restrict the number of allowed authorities. However, due to Thunderbird sharing its code basis with the Firefox browser, it is extremely complicated to load any files from disk into the Add-on. Especially files for which only the relative path is known. Therefore, after extensive investigations into this, it was decided that this would be too complicated and time intensive to implement. Furthermore, the advantages of this feature are rather slim, as organizations can simply take the plugins source code and manipulate the list directly in the code. Also updates to this authority list can be provided by updates of the whole Add-on.
When the user is satisfied with the report, he or she can send the report, which is done by clicking the send button in the upper left corner, just as it would be done in a regular e-mail compose window. For the sending of the report, the already configured e-mail server is used.
After the report has been sent the Thunderbird Add-on tags the reported e-mail as reported. In Thunderbird there are two basic ways to tag e-mails. The first way is to use Thunderbird's tagging mechanic. This will put a marker in Thunderbird's e-mail database with the tag text. Additionally, it will change the display of the e-mail in the Thunderbird UI by coloring the background differently. However, each tag is referenced by its own string, which means that no e-mail specific information can be stored in those tags without creating vast amounts of them.
The other tagging mechanism is to use X- mail headers. These headers can store tags with arbitrary text, but they are not displayed automatically. Therefore, to be visible they require another overlay which reads them and displays them. However their biggest disadvantage is that Thunderbird can not automatically search for these headers.

Since none of these two approaches is perfect, they are combined in the Thunderbird Add-on. Therefore, each reported e-mail is tagged with the Thunderbird tag to signal that it was reported already and a X- header is placed that shows to whom the mail was reported. Thereby it is possible to search for all reported e-mails as well as see to whom an e-mail was reported. This simplifies questions regarding such reports in later stages. Additionally it allows for the detection of duplicate reports, meaning a report is sent to the same authority twice. In case of a report duplication the Add-on warns the user, which can then act upon the warning.

### 4.3.3 Aggregator

The aggregator periodically checks the reporting mail box for new e-mails. When a new e-mail has arrived, it is first checked whether the e-mail is an X-ARF report or not, by looking at the *X-XARF* or *X-ARF* header. If the e-mail is an X-ARF report, then the message is taken by the aggregator and split into the three MIME parts. Afterwards the second MIME part can be structurally verified. This means that the aggregator can use the report schema to compare the information in the second MIME part with the information that should be included according to the schema. Through this validation, broken reports can be sorted out before they can cause problems. Additionally this structural validation makes attacks through fake reports more difficult, as the attacker is quite limited on what to include in the second MIME part, which is the only part of the report that is automatically processed. The aggregator then can use the information to cluster the reports. This clustering can for example be performed according to the report sender address, sending mail server IP address of a reported e-mail, or the subject of the reported e-mail. How the reports are clustered is up to the organizations internal report handling processes. The reason why the reports are clustered at all is to reduce the effort it takes an organization to analyze and act upon the reports. The example aggregator in this thesis clusters the report by the source fields in the report. However, in an operational context, it is easy to image the clustering being based on the similarity between reports instead of single fields. The report is then saved to disk in a file in a specific directory. This is done to allow experts to search through and analyze these reports with the tools they already use. This removes the necessity for experts to learn new processes.
During the analysis of the report, an expert decides whether the reported suspicious e-mail is an abusive e-mail and what type of e-mail abuse it is. The organization then can decide what countermeasures to take and provide feedback to the reporter. At last, the organization also has the option to share the report with external authorities. To share the report, the organization can anonymize the report and then transform it into a reporting format which the specific authority accepts. This allows the organization to utilize and support the additional protection of the external report accepting authorities, such as browser blacklists, while at the same time minimizing the risk of exposing sensitive information.

# 5 Evaluating the new Approach

In order to evaluate the new approach to end user e-mail abuse reporting, the novel approach has to be compared to old reporting processes as well as other possible ways to increase the number and quality of e-mail abuse reports.

The first major difference between the new approach and the old abuse reporting process is that in the new approach suspicious e-mails are reported instead of spam and phishing e-mails. These suspicious e-mails do not necessarily constitute any incidents as phishing e-mails or spam e-mails do. Therefore, it is possible that some legitimate e-mails are reported as suspicious as well. In a perfect world this constitutes a huge disadvantage, because report receivers have to classify the reports into real incidents and legitimate e-mails that were perceived as suspicious, which can increase the effort required to process these reports significantly. However, since many end users are not able to reliably classify e-mails correctly as legitimate, spam or phishing on their own, report receivers already have to validate each report and decide whether the e-mail really was a phishing or spam e-mail. Therefore, classifying an e-mail which was reported as suspicious will not be significantly more expensive than validating a spam, phishing or abusive e-mail report already is.

But suspicious e-mail reports have another disadvantage. Many clearing houses and authorities have specialized in one filed of e-mail abuse reporting. For example phishtank specialized in phishing reports. Therefore, these authorities might, at least initially, not be willing to receive suspicious e-mail reports. This would limit the authority pool significantly. This disadvantage can be overcome by simply finding one authority which takes suspicious e-mail reports and classifies the received reported e-mails. These newly classified e-mails can then be reported to the already established authorities. Such a classifying authority can either be found in a publicly available authority or be created organization internally.

The major advantage of suspicious e-mail reporting is the fact that end users do not have to classify e-mails into categories any more. This classification is extremely difficult for end users, as was shown in Section 2.1.5 and Section 3.7. Therefore, the end users were often unsure in their classification and might not report the e-mail in question. This can be completely remediated by allowing them to report suspicious e-mails, thereby potentially increasing the amount of e-mails that are reported significantly.

An alternative way to enable more people to report abusive e-mails as well as increase the quality of reports is to educate the end users, as discussed in Section 2.1.8. This leads to the users being more educated in detecting and reporting e-mails. However, this has been found to be extremely difficult and not 100% effective. Additionally, repeated training is rather expensive, while allowing end users to report suspicious e-mails only requires available resources in the security department. This can also be expensive, but suspicious e-mail reports provide even more information to the security department. For

example these reports can reveal that some legitimate internal e-mails are often perceived as suspicious. Thereby these e-mails might not be read by everybody who receives them. Furthermore, suspicious e-mail reports can be used to judge the employees ability to detect phishing e-mails and thereby indicate which individual potentially require more training. Moreover, suspicious e-mail reporting allows for direct feedback to the trainees and can thereby be used to increase the awareness even further. Thereby increasing the effectiveness of the training.

The next major difference between the standard e-mail abuse reporting process and the new approach is the fact, that the report generation is completely automated. This has the main advantage, that the report can be generated in mere seconds and does not require extensive amounts of time and manual actions, as before. Therefore, it reduces the discomfort of reporting abusive e-mails for end users drastically. Furthermore, the report will definitely contain a number of important information for processing the report. With reports that were manually generated, this was not the case as each reporter had to think for them selves which information to include and which not to include. This is especially problematic because end user usually do append the reported e-mail incorrectly, as they either only copy the content of the e-mail or forward the e-mail. In both cases the e-mail headers, which are often vital to the report validation and processing are lost.

Hence, the automatic report generation is a huge advantage. But at the same time this creates problems regarding the end user's privacy. the reason for this is that the automatically generated reports might contain sensitive information which the user might not want to share with external parties. Therefore, it is extremely important to allow the end users to anonymize certain parts of the report, which is also required by the new approach.

Nevertheless, the automatic generation of the reports allows for the establishment of a specific report format for suspicious e-mail reports. This brings the next main advantage of the new process, which is the ability to create reports that can be automatically processed. This alleviates the problems of handling larger volumes of reports. The usage of X-ARF in this case brings two other advantages with the new process. First and foremost, X-ARF can be used to report a diverse set of incidents. Therefore, it is possible to include the new suspicious e-mail reports into already existing incident handling processes which use X-ARF.

Additionally, X-ARF is extremely versatile. The first report part simply contains human readable text, which can be adapted to any incident report rather easily. The second part of the report contains a machine-readable and verifiable representation of the automatically parsed report information. Since X-ARF allows for the definition of report schemata, it is possible to design reports specifically catered to suspicious e-mail reports. These reports can clearly define which information are contained in this second report part. On the other hand, due to the already mentioned privacy concerns, the schema for suspicious e-mail reports has to contain either a lot of optional fields or be anonymizable. Since anonymization by end users can often lead to breaking certain format requirements, allowing for an anonymization of the information in the schema is a bad idea. Thus the schema has to contain the optional fields instead, which allow the end user to simple anonymize the other report parts and exclude the sensitive information form the information included in the second part. This however leads to the

problem, that different reports might include different amounts of optional information. This can not be prevented as there are no other options to ensure the reporters privacy in a generally valid fashion. But in an organization internal context these privacy concerns can play only minor roles. Therefore, organization internal reports can change these optional fields into required fields, which increases the amount of information that is definitely included in the report.
Last but not least the third X-ARF report part is perfect for including the reported e-mail, as it is defined as an RFC 2822 attachment to the report. Thus all e-mail headers remain intact.
Even though the report's content depends on the contained optional fields, the automatically generated reports will be of a consistent and in comparison to today's manually composed messages good, quality. This is a huge advantage of the new reporting process, as report receivers can rely on a certain report quality and thereby utilize more reports.
The last major difference regarding the old process is that the new approach explicitly allows and encourages organization internal suspicious e-mail reporting. This allows organizations to observe e-mail abuse directed at them and enables the organizations to enact counter measures. Furthermore, the internal reporting remediates the risk of information leakage through reports, which can be a major risk for organizations.
The only problem the new suspicious e-mail reporting process has is the fact that it relies on an reporting tool to be fully implemented. However, in Section 4.3.2 it could be shown that the development of such a reporting tool is possible. Thereby enabling the suspicious e-mail reporting process to be fully implemented and used.

# 6 Conclusion

In this thesis the reasons why end users do not report abusive e-mails were researched. In a conducted survey it was found that end users at large would like to report abusive e-mail, however most lack the knowledge about authorities to whom they can report. Furthermore, end users often have no experience with incident reporting. This poses a major problem for them as the original reporting process for abusive e-mails requires such knowledge and experience to be performed successfully. Moreover, end users also have problems identifying phishing e-mails reliably. Therefore, end users are often unsure about their classification of phishing e-mails. Thus, if they can just report spam and phishing e-mails, they might not report e-mails they are unsure about.

To combat all of these problems a new reporting process for e-mail abuse reporting is proposed in order to enable end users to report potentially abusive e-mails. The first difference between the old and the new reporting process is that while in the old process either a spam report or a phishing report was created, in the new process only suspicious e-mail reports are created. These suspicious e-mail reports can then be verified by experts who are better at differentiating between spam, phishing and legitimate e-mails. Therefore, no e-mail is reported under the wrong category and therefore dismissed.

Furthermore, the reporting process requires a tool to be present, which can generate a report and propose a report receiver with the press of a button. This makes the report generation for the end user incredibly easy. In the old process, the end user would have needed to look up an authority which would receive the report. Then he or she would need to create a report, in which the e-mail is included in a way that preserves the e-mail meta-data such as the headers. This whole ordeal is quite time consuming, complex and thereby error prone. In the new process all of these activities are performed automatically by the tool, which can be assumed to be a lot faster and more comfortable than the old process while also being less susceptible to mistakes.

This automatic report generation brings another advantage to the new process, which is a standardized report format. In the old reporting process, nothing is actually standardized. Therefore, reports were generated in a variety of different formats, most of them were even home grown ones. This prohibits the automatic analysis and processing of the reports. On the other hand the proposed process defines the X-ARF format as the reporting format. The X-ARF format is chosen because it is human and machine readable and verifiable. Therefore, the reports produced in the new process can all be pre-processed automatically as well as handled completely manually. Furthermore, it allows for a huge degree of flexibility in their reports, making the format also viable for other report types. This allows for incident reporting architectures which are based on the X-ARF format, in which e-mail abuse reports can then be easily integrated.

After the generation of the report but before it is actually sent the new reporting process requires that

the reporting user has to have the chance to inspect the generated report and anonymize it if necessary. This allows for the reporters privacy and organizations sensitive data to be protected.
To validate, that the described process can be implemented in a single reporting tool, an Add-on for the Thunderbird e-mail client was developed. This Add-on provides an easy to use tool to create and send e-mail abuse reports. Additionally, in this thesis a report schema for the X-ARF suspicious e-mail reports to external authorities is proposed. This reporting schema provides a basis set of information, which are included in reports of private individuals or organizations to public authorities who wish to receive reports. This set is chosen to be as small as possible, such that the privacy of the reporter is not harmed. If the reports are only reported internally, then the organizations can define their own schema to include exactly the information they would like in their internal reports. Therefore, organizations can include more information that are interesting to them in reports and get the most benefits out of their reports, while still being able to keep their sensitive information private.
Moreover, for small organizations, the amount of reports might prove to much to handle. Therefore, the report process allows for an aggregator to be used to collect, store and group reports together. As a prove of concept in this thesis such an aggregator was also developed. This aggregator can load all reports from a specified address and store them on disc.

## 6.1 Future Work

While this thesis proposes a new reporting process, a lot of work remains to be done. First and foremost, there needs to be a study conducted, researching the usability of the proposed e-mail abuse reporting process. This usability study can be performed using the developed Add-on, which would also result in improvement proposals for the tool. Thereby an example implementation for a reporting tool could be created.
Furthermore, similar tools for other e-mail clients need to be created. Only with tools for a variety of e-mail clients larger tests over multiple organizations with different internal organization structures are feasible. Those tests however are necessary to evaluate the real world e-mail abuse reporting behavior change created through the new process.
Furthermore, other ideas of additional features for such a reporting tool could be analyzed. One of these feature would be to provide assistance to end users regarding the detection of phishing e-mails. This could potentially be done by displaying warnings if an e-mail matches some phishing criteria. How many criteria should be met before a warning is displayed has to be researched according to user preference and the distribution of said characteristics in phishing mails.
Moreover, the aggregator could be extended to contain much more analytic capabilities. An example would be to compare the whois entries of different linked domains to attribute phishing e-mails to the respective phishing campaigns and thereby to a specific organization. But which analytic functionality is useful and which is not has to be part of another future research.
Additionally it should be researched whether the internal reporting process in organizations can be used to evaluate how effective an anti-phishing training has been. If this would be possible, it would allow organizations to train their members according to their capability to detect phishing instead of training

them according to the time to the last training. This would then have the capacity to decrease the costs as well as the quality of such training.

# 7 Appendix

The appendix of this thesis is mostly contained on the CD that is provided with the thesis. Therefore, a short overview of the content on the CD is provided here. Furthermore, the questions used in the performed survey can be found here.

## 7.1 CD Content

The content of the CD is structured as follows. In the root directory there is a PDF version of this thesis. Furthermore, the file *xarf-suspicious-email-schema-external.json* contains the schema discussed in Section 4.3.1. Additionally, there are two folders, *development* and *thesis*.
The *thesis* folder contains the latex source code necessary to recreate this thesis, including the used illustrations.
In the development folder, there are two sub-folders, *addon* and *aggregator*. The *addon* folder contains the source code of the Thunderbird Add-on. The source code for the, in Section 4.3.3, described aggregator is provided in the *aggregator* folder.

## 7.2 Questionnaire Questions

On the next pages the conducted survey is printed out.

**Please read the following statement with care. If you agree to take part in this study proceed by clicking the "Next" button.**

**Participant Information The purpose of this information sheet is to provide you with sufficient information so that you can then give your informed consent. It is thus very important that you read this document carefully, and raise any issues that you do not understand with the researcher. Name of Researcher: Adrian Metzner Name of Supervisor: Dieter Gollmann, Sven Uebelacker Institution of Study: Security in Distributed Applications, Hamburg University of Technology Project Subject: Phishing Questionnaire 1. What is the purpose of the project? The questionnaire poses questions regarding knowledge about phishing e-mails and incident reporting. They are part of my master thesis, in which I develop a tool enabling users to report phishing attacks. This enables the authorities to then act upon the information they received. The term "Phishing e-mail" describes e-mails which try to trick the recipient into giving the sender information or access to valuable resources. 2. Why have I been selected to take part and what are the criteria?**

**Participation is voluntary. Participants must be over the age of 18.**

**3. What will I have to do?**

**You will be asked to access the survey online and then provide informed consent by clicking on the "Next" button. The first section contains questions regarding your personality and behaviour in certain situations. The second part contains a link to a phishing test. After completing the test you are asked to answer questions regarding the results of the test. Afterwards, I will ask you a few questions regarding your experience with phishing e-mails in general. The following section will then pose a few questions regarding security incident reporting. The survey ends with a few questions about your preferences regarding reporting suspicious e-mails.**

**4. How will confidentiality be assured and who will have access to the information that I provide?**

**All data that you provide will be completely anonymous, as no names, email or IP addresses will be collected by the survey software ("limesurvey"). All data will be stored on a secure server located inside the Hamburg University of Technology (TUHH) and accessed and analyzed by the researchers only. The server is administered by TUHH staff (not by external survey providers).**

**5. If I require further information, who should I contact and how?**

**If you wish to find out any further information about the study or wish to know the results, you can contact the researcher by email at adrian.metzner@tuhh.de.**

**Participant Consent**

# Section A: Personality, Behavior, and Risk Taking

## A1. SeBIS

| | Never | Rarely | Sometimes | Often | Always |
|---|---|---|---|---|---|
| When I'm prompted about a software update, I install it right away. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I try to make sure that the programs I use are up-to-date. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I manually lock my computer screen when I step away from it. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I set my computer screen to automatically lock if I don't use it for a prolonged period of time. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I use a PIN or passcode to unlock my mobile phone. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I use a password/passcode to unlock my laptop or tablet. | ☐ | ☐ | ☐ | ☐ | ☐ |
| If I discover a security problem, I continue what I was doing because I assume someone else will fix it. | ☐ | ☐ | ☐ | ☐ | ☐ |
| When someone sends me a link, I open it without first verifying where it goes. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I verify that my antivirus software has been regularly updating itself. | ☐ | ☐ | ☐ | ☐ | ☐ |
| When browsing websites, I mouseover links to see where they go, before clicking them. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I know what website I'm visiting based on its look and feel, rather than by looking at the URL bar. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I do not change my passwords, unless I have to. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I use different passwords for different accounts that I have. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I do not include special characters in my password if it's not required. | ☐ | ☐ | ☐ | ☐ | ☐ |
| When I create a new online account, I try to use a password that goes beyond the site's minimum requirements. | ☐ | ☐ | ☐ | ☐ | ☐ |
| I submit information to websites without first verifying that it will be sent securely (e.g., SSL, "https://", a lock icon). | ☐ | ☐ | ☐ | ☐ | ☐ |

## A2. DOSPERT-RT

**For each of the following statements, please indicate the likelihood that you would engage in the described activity or behavior if you were to find yourself in that situation.**

| | Extremely Unlikely | Moderately Unlikely | Somewhat Unlikely | Not Sure | Somewhat Likely | Moderately Likely | Extremely Likely |
|---|---|---|---|---|---|---|---|
| Admitting that your tastes are different from those of a friend. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Disagreeing with an authority figure on a major issue. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Extremely Unlikely | Moderately Unlikely | Somewhat Unlikely | Not Sure | Somewhat Likely | Moderately Likely | Extremely Likely |
|---|---|---|---|---|---|---|---|
| Choosing a career that you truly enjoy over a more secure one. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Speaking your mind about an unpopular issue in a meeting at work. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Moving to a city far away from your extended family. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Starting a new career in your mid-thirties. | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

## A3. BIS-Brief

| | rarely/never | occasionally | often | almost always/always |
|---|---|---|---|---|
| I plan tasks carefully. | ☐ | ☐ | ☐ | ☐ |
| I do things without thinking. | ☐ | ☐ | ☐ | ☐ |
| I am self controlled. | ☐ | ☐ | ☐ | ☐ |
| I am a careful thinker. | ☐ | ☐ | ☐ | ☐ |
| I say things without thinking. | ☐ | ☐ | ☐ | ☐ |
| I act on the spur of the moment. | ☐ | ☐ | ☐ | ☐ |
| I don't "pay attention". | ☐ | ☐ | ☐ | ☐ |
| I concentrate easily. | ☐ | ☐ | ☐ | ☐ |

## A4. I see myself as:

| | disagree strongly | disagree moderately | disagree a little | neither agree nor disagree | agree a little | agree moderately | agree strongly |
|---|---|---|---|---|---|---|---|
| extraverted, enthusiastic | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| critical, quarrelsome | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| dependable, self-disciplined | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| anxious, easily upset | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| open to new experiences, complex | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| reserved, quiet | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| sympathetic, warm | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| disorganized, careless | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| calm, emotionally stable | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | disagree strongly | disagree moderately | disagree a little | neither agree nor disagree | agree a little | agree moderately | agree strongly |
|---|---|---|---|---|---|---|---|
| conventional, uncreative | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

# Section B: Phishing Test Experiment

Please go to the following link

**https://www.sonicwall.com/phishing/phishing-quiz-question.aspx**

and complete the phishing test there. After completing the test please fill in the questions in this section regarding your results in the test.

**B1. Did you understand your task as well as the hints given in the phishing test?**

- ☐ Yes
- ☐ No

**B2. Test results:**

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| How many e-mails could you classify correctly? | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| How many phishing e-mails did you not recognize as such? | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| How many valid e-mails did you classify as phishing e-mails? | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**B3. Did you learn new techniques to spot phishing e-mails?**

- ☐ Yes
- ☐ No
- ☐ Maybe

**B4. How confident were you during the phishing test, that you categorized the presented e-mail correctly?**

- ☐ I was absolutely sure that I always chose the correct answer
- ☐ On 1 or 2 e-mails I was not sure whether these e-mails were valid e-mails or phishing e-mails
- ☐ On more then 2 occasions I was unsure whether the e-mail was legitimate or not

**B5. When presented with the e-mails you were unsure about, would you have liked to be able to ask for a second opinion?**

- ☐ Yes
- ☐ No

**B6.** **Why wouldn't you have asked someone else for his opinion?**

## Section C: Experience with Phishing

**C1.** **In your office, do you think that phishing attacks are a risk ...**

| | strongly agree | agree | disagree | strongly disagree |
|---|---|---|---|---|
| for your organization | ☐ | ☐ | ☐ | ☐ |
| for your privacy | ☐ | ☐ | ☐ | ☐ |

**C2.** **Have you ever received an e-mail that you identified as a phishing e-mail?**

- ☐ Yes
- ☐ No

**C3.** **Do you know someone who trusted a phishing mail?**

- ☐ Yes
- ☐ No

**C4.** **Have you ever trusted a phishing mail directed at your professional account?**

- ☐ Yes and it had negative consequences for both the company and myself
- ☐ Yes and it had negative consequences for the company but not for myself
- ☐ Yes and it had negative consequences for me but not for the company
- ☐ Yes but it had no negative consequences for me or the company
- ☐ No
- ☐ Not sure

**C5.** **Have you ever trusted a phishing mail directed at your private e-mail account?**

- ☐ Yes and it had negative consequences
- ☐ Yes but it had no negative consequences
- ☐ No
- ☐ Not sure

**C6. Who should be responsible for mitigating the risk of a successful phishing attack against your professional e-mail account?**

- ☐ the organization/company through administrators or the security department
- ☐ each employee for himself
- ☐ both

**C7. Do you agree with the following statements?**

| | strongly agree | agree | disagree | strongly disagree |
|---|---|---|---|---|
| My anti-virus software, spam filters and firewalls prevent me from becoming a victim of phishing attacks | ☐ | ☐ | ☐ | ☐ |
| Warnings about phishing mails directed at other employees help me to identify potential phishing attempts targeted at myself | ☐ | ☐ | ☐ | ☐ |

# Section D: Reporting

**D1. Have you ever reported a security related incident to an authority inside or outside of your organization?**

- ☐ Yes
- ☐ No

**D2. Have you ever reported a security related incident that occurred privately to an outside authority?**

- ☐ Yes
- ☐ No

**D3. Do you know authorities to whom security incidents can be reported?**

- ☐ Yes
- ☐ No

**D4. To whom would you report security incidents?**

*Please enter the name of the authorities you would report incidents to. For example : "security department, Management"*

# Section E: Reporting tool

**E1. Do you agree with the following statements?**

| | strongly agree | agree | disagree | strongly disagree |
|---|---|---|---|---|
| I want to report suspicious e-mails sent to my private e-mail account to an outside authority to protect other individuals from potential phishing attacks | ☐ | ☐ | ☐ | ☐ |
| I want to report suspicious e-mails sent to my professional e-mail account to an internal authority to to protect the organization from potential phishing attacks | ☐ | ☐ | ☐ | ☐ |

**E2. Imagine you reported a suspicious e-mail. In which way would you like to receive feedback?**

- ☐ A personalized message informing me whether it was a phishing attempt or not
- ☐ Weekly/Monthly organization wide information regarding the phishing schemes which are run against the organization
- ☐ Only warnings from especially dangerous phishing attacks
- ☐ No feedback at all

**E3. How fast would you like to receive feedback?**

- ☐ Within an hour
- ☐ At the same day
- ☐ Within a week
- ☐ Never

**Social Engineering Questionnaire**

**Name of Researcher: Adrian Metzner Name of Supervisor: Dieter Gollmann, Sven Uebelacker Institution of Study: Security in Distributed Applications, Hamburg University of Technology Project Subject: Social Engineering Questionnaire 1. What is the purpose of the project? The questionnaire poses questions regarding knowledge about phishing e-mails and incident reporting. They are part of my master thesis, in which I develop a tool enabling users to report phishing attacks. This enables the authorities to then act upon the information they received. The term "Phishing e-mail" describes e-mails which try to trick the recipient into giving the sender information or access to valuable resources. 2. How will I find out about the results?**

**If you would like to know the results of this study, please contact Adrian Metzner at adrian.metzner@tuhh.de. You will be able to request feedback about the results from the researcher once the data has been analyzed.**

**3. Have I been misled in any way during the project?**

**No, there is no deception involved in this survey.**

**4. How will the collected data be used?**

**The data collected in this study will be used to identify difficulties regarding the detection of phishing attempts. It will also be used to capture the requirement users have for a reporting tool. The results will be used in my master thesis. Results may also be published in scientific journals or presented at conferences. Informa- tion and data gathered during this research study will only be available to the researchers working on this study. Should the research be presented or published in any form, all data will be anonymous (i.e., your personal information or data will not be identifiable). If the research is of publishable quality, the data may be kept for up to 5 years before being destroyed. During that time the data may be used by the researchers only for purposes appropriate to the research question, but at no point will your personal information or data be revealed.**

**Thank you for your participation. Please close this browser window to exit the survey.**

**(the phrasing of this page is mostly based on the Job Adverts as a Recruitment Tool survey of Edinburgh Napier University.)**

# Acknowledgements

The creation of this thesis would not have possible without the support of many different people and organizations. However, some people deserve to be mentioned explicitly.
First and foremost, I would like to thank Sven Übelacker, who supported me greatly by providing ideas and feedback throughout the thesis. Furthermore, he provided helpful feedback during the creation of the survey as well as the development of the plugin and the aggregator.
Moreover, I would like to thank the DFN-CERT, which provided counseling regarding X-ARF and e-mail abuse reporting. Additionally they were able to provide the viewpoint of a report receiving authority, which I would have been missing without them. From the DFN-CERT team I would like to thank specifically Dr.-Ing. Christian Keil and Youssef Rebahi-Gilbert who took time out of their busy schedules to listen to my ideas and correcting them when I was wrong.
I would also like to thank Margret Ford, who assisted me with the creation of the survey by ensuring, that the survey retained a certain linguistic quality and unambiguity.
I also have to thank the team of the *Rechenzentrum* (Computing center) of the technical university hamburg, who provided me with the perspective of an internal organization which is experiencing the problems end users face when reporting abusive e-mails.

# List of Figures

# Listings

# List of Tables

# Bibliography


[1] abusix GmbH. x-arf: Network abuse reporting 2.0. `http://x-arf.org`. Accessed: 2017-01-02.

[2] Tiago A Almeida, Jurandy Almeida, and Akebo Yamakami. Spam filtering: how the dimensionality reduction affects the accuracy of naive bayes classifiers. *Journal of Internet Services and Applications*, 1(3):183–200, 2011.

[3] Tiago A Almeida and Akebo Yamakami. Content-based spam filtering. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–7. IEEE, 2010.

[4] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 82:69 – 82, 2015.

[5] Oren Ben-Kiki, Clark Evans, and Brian Ingerson. Yaml ain't markup language (yamlâĎć) version 1.2. *yaml.org, Tech. Rep*, 2005.

[6] Tim Berners-Lee, Roy Fielding, and H Frystyk. Rfc 1945: Hypertext transfer protocolâĂŤhttp/1.0, may 1996. Technical report, RFC, 2005.

[7] Ann-Renée Blais and Elke U Weber. A domain-specific risk-taking (dospert) scale for adult populations. *Judgment and Decision Making, Vol. 1, No. 1*, 2006.

[8] itnews Brett Winterford. Epsilon breach used four-month-old attack. `https://www.itnews.com.au/news/epsilon-breach-used-four-month-old-attack-253712`. Accessed: 2017-03-08.

[9] Nevil Brownlee. Expectations for computer security incident response. Technical report, RFC, 1998.

[10] Guy Bruneau. It is tax season - watch out for suspicious attachment. `https://isc.sans.edu/forums/diary/It+is+Tax+Season+Watch+out+for+Suspicious+Attachment/22117/`. Accessed: 2017-04-18.

[11] Finn Brunton. *Spam: A shadow history of the Internet*. MIT Press, 2013.

[12] Laurence A Canter and Martha S Siegel. *How to Make a Fortune on the Information Superhighway: What the Internet Is, how It Works*. HarperTrade, 1995.

[13] Deanna D Caputo, Shari Lawrence Pfleeger, Jesse D Freeman, and M Eric Johnson. Going spear phishing: Exploring embedded training and awareness. *IEEE Security & Privacy*, 12(1):28–38, 2014.

[14] Madhusudhanan Chandrasekaran, Krishnan Narayanan, and Shambhu Upadhyaya. Phishing email detection based on structural properties. In *NYS Cyber Security Conference*, pages 1–7, 2006.

[15] Robert B. Cialdini and Noah J. Goldstein. The science and practice of persuasion. *Cornell Hotel and Restaurant Administration Quarterly*, 43(2):40–50, 2002.

[16] Robert B Cialdini, Joyce E Vincent, Stephen K Lewis, Jose Catalan, Diane Wheeler, and Betty Lee Darby. Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique. *Journal of personality and Social Psychology*, 31(2):206, 1975.

[17] Colgate-Palmolive Company. Colgate totalÂő toothpaste the number 1 trusted choice of dental professionals. `http://www.colgateprofessional.com/health-benefits/colgate-total-triclosan`. Accessed: 2017-03-17.

[18] Lorrie Faith Cranor and Brian A. LaMacchia. Spam! *Commun. ACM*, 41(8):74–83, August 1998.

[19] Dave Crocker. Mailbox names for common services, roles and functions. 1997.

[20] Roman Danyliw, Jan Meijer, and Yuri Demchenko. The incident object description exchange format. Technical report, ietf, 2007.

[21] Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590. ACM, 2006.

[22] Brad Duncan. Brazilian malspam sends autoit-based malware. `https://isc.sans.edu/forums/diary/Brazilian+malspam+sends+Autoitbased+malware/22081/`. Accessed: 2017-04-18.

[23] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2873–2882. ACM, 2015.

[24] Clarc Evans. Yaml ain't markup language. `yaml.org`. Accessed: 2017-02-17.

[25] Nathaniel Joseph Evans. *Information technology social engineering: an academic definition and study of social engineering-analyzing the human firewall*. PhD thesis, Citeseer, 2009.

[26] Ana Ferreira, Lynne Coventry, and Gabriele Lenzini. Principles of persuasion in social engineering and their use in phishing. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 36–47. Springer, 2015.

[27] Jerry Finn. A survey of online harassment at a university campus. *Journal of Interpersonal violence*, 19(4):468–483, 2004.

[28] Center for Decision Sciences | Columbia Business School. Dospert scale. `http://dospert.org`. Accessed: 2017-01-20.

[29] International Society for Research on Impulsivity. Bis 11. `http://www.impulsivity.org/measurement/bis11`. Accessed: 2017-01-20.

[30] H Clayton Foushee. Dyads and triads at 35,000 feet: Factors affecting group process and aircrew performance. *American Psychologist*, 39(8):885, 1984.

[31] Jonathan L Freedman and Scott C Fraser. Compliance without pressure: the foot-in-the-door technique. *Journal of personality and social psychology*, 4(2):195, 1966.

[32] Jonathan D Frieden and Sean Patrick Roche. E-commerce: Legal issues of the online retailer in virginia. *Rich. JL & Tech.*, 13:1, 2006.

[33] Samuel D Gosling, Peter J Rentfrow, and William B Swann. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

[34] David Gragg. A multi-level defense against social engineering. *SANS Reading Room, March*, 13, 2003.

[35] Anti-Phishing Working Group et al. Phishing archive, 2005.

[36] Anti-Phishing Working Group et al. Phishing activity trends report 1st - 3rd quarters 2015. *Anti-Phishing Working Group*, 2015.

[37] Anti-Phishing Working Group et al. Phishing activity trends report 4th quarter 2016. *Anti-Phishing Working Group*, 2016.

[38] J Hong. Why have there been so many security breaches recently? blog@ cacm (apr. 27, 2011). Accessed: 2017-03-08.

[39] Alice Hutchings, Richard Clayton, and Ross Anderson. Taking down websites to prevent crime. In *Electronic Crime Research (eCrime), 2016 APWG Symposium on*, pages 1–10. IEEE, 2016.

[40] SonicWall Inc. Sonicwall phishing iq test. `https://www.sonicwall.com/phishing/phishing-quiz-question.aspx`. Accessed: 2017-01-20.

[41] Tom N Jagatic, Nathaniel A Johnson, Markus Jakobsson, and Filippo Menczer. Social phishing. *Communications of the ACM*, 50(10):94–100, 2007.

[42] Graham Klyne and Chris Newman. Date and time on the internet: Timestamps. Technical report, RFC, 2002.

[43] Erka Koivunen. "Why Wasn't I Notified?": Information Security Incident Reporting Demystified. In *NordSec*, 2010.

[44] Ponnurangam Kumaraguru, Yong Rhee, Steve Sheng, Sharique Hasan, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Getting users to pay attention to anti-phishing education: evaluation of retention and transfer. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pages 70–81. ACM, 2007.

[45] Daniel Lowd and Christopher Meek. Good Word Attacks on Statistical Spam Filters. In *CEAS*, 2005.

[46] Christian Ludl, Sean McAllister, Engin Kirda, and Christopher Kruegel. On the effectiveness of techniques to detect phishing sites. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 20–39. Springer, 2007.

[47] Steve Mansfield-Devine. Ransomware: taking businesses hostage. *Network Security*, 2016(10):8–17, 2016.

[48] Robert R McCrae and Paul T Costa Jr. A five-factor theory of personality. *Handbook of personality: Theory and research*, 2:139–153, 1999.

[49] Kevin D Mitnick and William L Simon. *The art of deception: Controlling the human element of security*. John Wiley & Sons, 2011.

[50] JG Mohebzada, A El Zarka, Arsalan H BHojani, and Ali Darwish. Phishing in a university community: Two large scale phishing experiments. In *Innovations in Information Technology (IIT), 2012 International Conference on*, pages 249–254. IEEE, 2012.

[51] Antonio Nappa, M Zubair Rafique, and Juan Caballero. Driving in the cloud: An analysis of drive-by download operations and abuse reporting. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 1–20. Springer, 2013.

[52] Cleber K. Olivo, Altair O. Santin, and Luiz S. Oliveira. Obtaining the threat model for e-mail phishing. *Applied Soft Computing*, 13(12):4841 – 4848, 2013.

[53] OpenDNS. Phishtank. `https://www.phishtank.com/`. Accessed: 2017-01-28.

[54] Bimal Parmar. Protecting against spear-phishing. *Computer Fraud & Security*, 2012(1):8–11, 2012.

[55] Jonathan Postel. Rfc 821: Simple mail tranfer protocol. *Internet Engineering Task Force*, 1982.

[56] Anirudh Ramachandran, Anirban Dasgupta, Nick Feamster, and Kilian Weinberger. Spam or ham?: characterizing and detecting fraudulent not spam reports in web mail systems. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, pages 210–219. ACM, 2011.

[57] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.

[58] Justin M Rao and David H Reiley. The economics of spam. *Journal of Economic Perspectives*, 26(3):87 – 110, 2012.

[59] Dennis T Regan. Effects of a favor and liking on compliance. *Journal of Experimental Social Psychology*, 7(6):627–639, 1971.

[60] Bruce Rind and Prashant Bordia. Effect on restaurant tipping of male and female servers drawing a happy, smiling face on the backs of customers' checks. *Journal of Applied Social Psychology*, 26(3):218–225, 1996.

[61] Germany RUS-CERT, University of Stuttgart. Caif - common announcement interchange format. `http://www.caif.info/`. Accessed: 2017-02-17.

[62] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, pages 98–105, 1998.

[63] Jamison W Scheeres. Establishing the human firewall: reducing an individual's vulnerability to social engineering attacks. Technical report, DTIC Document, 2008.

[64] P Wesley Schultz. Changing behavior with normative feedback interventions: A field experiment on curbside recycling. *Basic and applied social psychology*, 21(1):25–36, 1999.

[65] Yakov Shafranovich, Murray S Kucherawy, and John Levine. An extensible format for email feedback reports. Technical report, ietf, 2010.

[66] Richard Shay, Iulia Ion, Robert W Reeder, and Sunny Consolvo. My religious aunt asked why i was trying to sell her viagra: experiences with account hijacking. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2657–2666. ACM, 2014.

[67] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie 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*, pages 373–382. ACM, 2010.

[68] Verizon Enterprise Solutions. 2016 data breach investigations report. *verizon. com*, 2016.

[69] Matthew S Stanford, Charles W Mathias, Donald M Dougherty, Sarah L Lake, Nathaniel E Anderson, and Jim H Patton. Fifty years of the barratt impulsiveness scale: An update and review. *Personality and individual differences*, 47(5):385–395, 2009.

[70] Stuart Staniford-Chen, Brian Tung, Dan Schnackenberg, et al. The common intrusion detection framework (cidf). http://gost.isi.edu/cidf/papers/cidf-isw.txt, 1998.

[71] Lynne Steinberg, Carla Sharp, Matthew S Stanford, and Andra Teten Tharp. New tricks for an old measure: The development of the barratt impulsiveness scale–brief (bis-brief). *Psychological assessment*, 25(1):216, 2013.

[72] Jessica Steinberger, Anna Sperotto, Mario Golling, and Harald Baier. How to exchange security events? overview and evaluation of formats and protocols. In *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, pages 261–269. IEEE, 2015.

[73] Stefan KÃűpsell Stephan Escher. DurchfÃijhrung eines integrierten anti-phishing-trainings. *DFN Konferenz*, 2016.

[74] Don Stikvoort. Istlp - information sharing traffic light protocol. Technical report, NISCC (UK), 2009.

[75] Robert Strahan and Kathleen Carrese Gerbasi. Short, homogeneous versions of the marlow-crowne social desirability scale. *Journal of clinical psychology*, 28(2):191–193, 1972.

[76] Olivier Thonnard and Marc Dacier. A strategic analysis of spam botnets operations. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, pages 162–171. ACM, 2011.

[77] The New York Times. Citibank statement on phishing email. http://markets.on.nytimes.com/research/stocks/news/press_release.asp?docTag=201703170435BIZWIRE_USPRX____BW5219&feedID=600&press_symbol=68598. Accessed: 2017-04-18.

[78] The New York Times. Hackers use new tactic at austrian hotel: Locking the doors. https://www.nytimes.com/2017/01/30/world/europe/hotel-austria-bitcoin-ransom.html. Accessed: 2017-04-18.

[79] Sven Uebelacker. Visualisation of x-arf messages. https://camo.githubusercontent.com/4b629fe8b87a1313c4471a5848026be4f77b8ff7/68747470733a2f2f7261772e6769746875622e636f6d2f786172662f786172662d737065636966696361746 2016.

[80] Sven Uebelacker and Susanne Quiel. The social engineering personality framework. In *2014 Workshop on Socio-Technical Aspects in Security and Trust*, pages 24–30. IEEE, 2014.

[81] Steven J Vaughan-Nichols. Saving private e-mail. *IEEE Spectrum*, 40(8):40–44, 2003.