An Analysis of Fusing Advanced Malware Email Protection Logs, Malware Intelligence and Active Directory Attributes as an Instrument for Threat Intelligence

Submitted in partial fulfilment
of the requirements of the degree of

Master of Science

of Rhodes University

Japie Vermeulen

Grahamstown, South Africa
17th March 2018

Abstract

After more than four decades email is still the most widely used electronic communication medium today. This electronic communication medium has evolved into an electronic weapon of choice for cyber criminals ranging from the novice to the elite. As cyber criminals evolve with tools, tactics and procedures, so too are technology vendors coming forward with a variety of advanced malware protection systems. However, even if an organization adopts such a system, there is still the daily challenge of interpreting the log data and understanding the type of malicious email attack, including who the target was and what the payload was.

This research examines a six month data set obtained from an advanced malware email protection system from a bank in South Africa. Extensive data fusion techniques are used to provide deeper insight into the data by blending these with malware intelligence and business context. The primary data set is fused with malware intelligence to identify the different malware families associated with the samples. Active Directory attributes such as the business cluster, department and job title of users targeted by malware are also fused into the combined data.

This study provides insight into malware attacks experienced in the South African financial services sector. For example, most of the malware samples identified belonged to different types of ransomware families distributed by known botnets. However, indicators of targeted attacks were observed based on particular employees targeted with exploit code and specific strains of malware. Furthermore, a short time span between newly discovered vulnerabilities and the use of malicious code to exploit such vulnerabilities through email were observed in this study. The fused data set provided the context to answer the "who", "what", "where" and "when". The proposed methodology can be applied to any organization to provide insight into the malware threats identified by advanced malware email protection systems. In addition, the fused data set provides threat intelligence that could be used to strengthen the cyber defences of an organization against cyber threats.

Acknowledgements

First and foremost, I would like to thank my amazing family for their love, support and understanding of the family time sacrificed during the past two years to complete my studies. I would also like to thank Prof. Karen Bradshaw my supervisor for all the proof reading, guidance and correcting my mistakes. I also would like to extend a sincere thanks to Prof. Barry Irwin for the first year of guidance and support. Finally, I would like to thank my employer for granting me the time off, often in very short notice to work on my thesis and all the assignments.

ACM Computing Classification System Classification

Thesis classification under the ACM Computing Classification System[1] (2012 version valid through 2017):

- **Security and privacy~Malware and its mitigation**

- *Mathematics of computing~Exploratory data analysis*

[1] https://dl.acm.org/ccs/ccs_flat.cfm

Contents

List of Figures

List of Tables

List of Code Listings

Chapter 1

Introduction

"An amazing thing, the human brain. Capable of understanding incredibly complex and intricate concepts. Yet at times unable to recognize the obvious and simple." – Jay Abraham

1.1 Context of Research

Employees with an inadequate understanding of the dangers associated with the Internet, pose a significant information security risk to the organizations they are employed in. Furthermore, with email being the most widely used medium for malware distribution, organizations face a constant threat of unsuspecting employees opening malicious emails (Dhamija & Hearst, 2006). To mitigate this risk, an organization can deploy an Advanced Malware Protection (AMP) solution to prevent malicious emails from entering the organization's network as part of a layered defense architecture (Cisco, 2015).

However, for an organization to be aware of any malicious email campaigns targeting them, correlation and analysis of the required logs generated by such a solution are required (Lumeta, 2015). Furthermore, by fusing additional contextual data log sources together, an organization can derive intelligent information from the combined output of different data log sources (Informatica, 2013). For example, the use of Active Directory (AD) data in conjunction with the AMP log data can provide more context such as the business unit in which a user resides or the job role of a particular user (Microsoft, 2016). Therefore, the use of security analytics (Informatica, 2013) together with Business Intelligence (BI) methods for the analysis of such fused logs, can help identify hidden patterns of malicious activity never seen before in an organization (Hoppe *et al.*, 2009) as well as aid in the decision making progress within an organization (Chen & Storey, 2012).

Such data analysis techniques can also be applied to real-time monitoring as part of continuous cyber situational awareness monitoring (Lumeta, 2015). Should any information security risks be uncovered based on previously unseen malicious activity patterns being identified by such analysis methods, an organization could respond in a timely manner to address any risk identified (Hoppe *et al.*, 2009).

1.2 Motivation for this Research

While in full-time employment at a South African financial institution, the researcher often overheard questions regarding the benefits of an AMP solution such as:

- "What exactly does our expensive advanced malware email protection solution protect us against?"
- "Who is being targeted in our organization?"

This research set out to gain insight into malware driven email attacks on the financial industry in South Africa as part of the "Knowledge Discovery of Digital Data" (KDD) to identify the level of threat intelligence that can be obtained from the data. Knowledge discovery is the process of uncovering intelligent information that is of value and usable to an organization (Fayyad *et al.*, 1996). In addition, we hypothesize that fusing data from a variety of sources can provide more meaningful knowledge discovery.

1.3 Research Questions

In the process of proving our hypothesis, the following research questions were posed:

- To what extent can the fusion of AMP logs, malware intelligence and AD user attributes be used as an instrument for threat intelligence?
- To what extent can the fusion of AMP logs, malware intelligence and AD user attributes be used in a cyber security awareness program?
- What type of threat intelligence can be derived by the fusion of AMP logs with AD user attributes in an organization?
- Can the information derived from the above research questions be factored into a cyber security dashboard for executive management?

1.4 Research Objectives

The primary objective of this research project is to create a fused log system that can be used to provide actionable threat intelligence to an organization. To achieve this primary objective, the following secondary objectives are defined:

1. Analyse a variety of types of individual log data to ascertain what information is available from each.
2. Utilize a suite of tools that can be used to combine various types of log data into a fused data set.
3. Analyse the fused log data to ascertain the kind of information available.
4. Determine the usefulness of the information obtained for threat intelligence.

1.5 Expected Contributions of the Research

In Ponemon's research report "Big data analytics in Cyber Defense" (Ponemon Institute LLC, 2013) the researchers found that the analysis of large data sets obtained from security technologies provided organizations deeper insight to combat cyber threats. In addition, eighty-two percent of the respondents wanted the analysis of large security data sets combined with malware data. Teymourlouei & Jackson (2017) found that the analyses of large security data sets provided a quick solution to identify anomalies and advanced attack vectors. Hutchins *et al.* (2011) emphasized the importance of advanced malware data analysis as an instrument for intelligent network defense.

However, Fan *et al.* (2014) discussed some of the challenges posed in the processing and analysis of large data sets. For example, large data sets require significantly more computational processing power than normal data sets. In addition, large data sets tend to have a higher noise factor to work through before deriving intelligent data that is of value to the business.

This research provides a methodology that could be applied in organizations as required to derive threat intelligence from internal log data sources. Chismon & Ruks (2015) discussed the importance of determining the data requirements that would produce an intelligent data set that could be used as a source of threat intelligence.

1.6 Approach

A script was required to extract the required data from the different data log sources and write the output to file. The extracted data was fused with malware intelligence and AD attributes, thereby adding two layers of context to the data for further analysis and finally BI analysis methods were applied to the fused data to perform a knowledge discovery on the fused data set. Our approach consisted of six steps as set out below.

1. Obtain primary data set from the bank based on time frame provided.
2. Create scripts to extract the required malware intelligence and business context.
3. Check data consistency and clean data.
4. Anonomize data.
5. Fuse the primary data set by joining the extracted malware intelligence and business context data.
6. Apply BI analysis methods to the fused data.

1.7 Limitations of the Research

The researcher is aware of research studies that have been done on the effectiveness of advanced malware protection systems. However, a large portion of this research focuses on public and open-source based sandboxes. The data set used was obtained from a commercial sandbox that ranked as a leader in the Gartner[1] quadrant.

A further limitation of the study, is that only email attacks utilizing malicious attachments identified by the advanced malware protection system were studied. Furthermore, the data utilized for this study were specific to one of the five largest banks in South Africa.

1.8 Thesis Structure

The remainder of the thesis is arranged as follows:

- Chapter 2 provides essential background information on malware, phishing, log analysis, data fusion, threat intelligence and related research.
- Chapter 3 describes the research methodology, the collection and processing of the data sets and the data fusion process.
- Chapter 4 presents the results of the individual data set analyses. A comparison of the results from a South African perspective and an international data set is also provided and discussed.
- Chapter 5 presents the results of the fused data set analysis. The results in this chapter are discussed in the context of seven threat scenarios.
- Chapter 6 discusses the usefulness of the fused data set results by first contextualizing the limitation of each individual data set. This is followed by a discussion of how the results of the fused data set feeds into the four sub areas of threat intelligence discussed in Chapter 2.
- Chapter 7 concludes the research and suggests possible future research.

[1] https://www.gartner.com

Appendix A provides additional information about this research project in the form of a link to a Github[2] repository containing the unique list of SHA256 cryptographic identifiers obtained from the malware samples used in this research.

[2] https://github.com/

Chapter 2

Background Concepts and Related Work

"Reading is equivalent to thinking with someone else's head instead of with one's own." –Arthur Schopenhauer

In order to satisfy our research goal and the body of knowledge to which this research will contribute, some essential background knowledge is required. This chapter covers an overview of malware and phishing in particular, including defences against phishing. This is followed by a discussion on log data analysis and the fusion of a variety of log types. Different types of threat intelligence and the threat intelligence life cycle are presented next. Finally, related research is discussed.

2.1 Malware

"People's computers are not getting more secure. They're getting more infected with viruses. They're getting more under the control of malware." –Avi Rubin

Today malware is an umbrella term associated with almost any type of malicious software code used by cyber criminals (Rieck *et al.*, 2008). Over the last 10 years, malware has evolved from the pestilent auto-run[1] malware to

[1] http://www.howtogeek.com/203522/how-autorun-malware-became-a-problem-on-windows-and-how-it-was-mostly-fixed/

sophisticated malicious code used in attacks on nation states (Bencsáth *et al.*, 2012). Furthermore, malware has been attributed to the destruction of data on tens of thousands of machines, as well as halting operations at multi-billion dollar corporations (Bronk & Tikk-Ringas, 2013). In the same way, cyber criminals have developed different strains of sophisticated financial malware targeting banks and retailers globally with reported losses netting millions in local currency (Kaspersky Lab, 2015a; Symantec, 2014).

The analysis of malware therefore, forms an important part of understanding the type of malicious code being dealt with and the potential impact to an organization (Solutionary, 2005). In the following sections we discuss the identification of malware and the different malware categories and malware families.

2.1.1 Identification of malware

The first step in the identification of malware is to assign a cryptographic label to the malware sample that can be used as a cryptographic identifier (Kendall & McMillan, 2007). Such a label can be used to obtain more information about the malware without wasting countless hours of searching. The cryptographic identifier can be used to check online repositories for any known information about the particular malware and can also be used as an integrity check point during the analysis of malware. For example, if the original software code of the malware were to change, the cryptographic identifier would change as well (Kendall & McMillan, 2007).

2.1.2 Malware categories

Often when malware is discovered in an organization, it is important to understand to which category the malware belongs (Solutionary, 2005). For example, a piece of malware could belong to the "Worm" category, thereby containing malicious code that could spread through the network exploiting network vulnerabilities (DuPaul, 2012). Malware could also contain "Backdoor" functionality that would allow an attacker remote access into a network through a covert channel (Hardikar, 2008). This type of malware could be used to deploy additional malware, maintain persistence or steal information from an organization (Hardikar, 2008).

Nevertheless, many different types of malware categories exist providing different functionality as shown in Table 2.1 (Hardikar, 2008; DuPaul, 2012).

Table 2.1: The different malware categories

Malware Category	Description
Virus	Malware that copies itself and spreads to other computers
Worm	Malware that can exploit vulnerabilities and spread over the network
Backdoor	Malware that opens a covert channel to attackers on a system
Trojan	Malware that hides in a legitimate program; also used to steal information
Ransomware	Malware that renders data unusable until an amount is paid to an attacker
Keylogger	Malware that logs all keystrokes on a system
Rootkit	Malware that uses a specific method to hide in the operating system
Adware	A type of malware that displays advertising banners whilst using the Internet
Spyware	Malware used to spy on a target and steal information

2.1.3 Malware families

Although the different malware categories offer a glimpse of the functionality provided by a particular piece of malware (Hardikar, 2008), the behavioural patterns and code similarities of malware can be used to attribute the code to a particular malware family (Rieck *et al.*, 2008). For example, malware can behave in a specific way to interact with a system in a particular programmatic order. Furthermore, malware could contain static code similarities. Therefore, dynamic and/or static analysis of malicious code is used in the classification of malware and its attribution to a particular malware family (Rieck *et al.*, 2008).

The identification of a particular malware family could sometimes be attributed to a particular malware distribution network or cyber criminals known to make use of a particular type of malware (Symantec, 2015, 2017; Bartholomew & Guerrero-Saade, 2016). In July 2010, Stuxnet Lee (2012), the first of a kind of malware that was designed to cause physical damage to critical infrastructure, was discovered by security researchers from Belarus. However, security researchers from Symantec traced Stuxnet back to November 2005 (McDonald *et al.*, 2013). The Stuxnet worm was specifically designed to alter the code on Programmable Logic Controllers found in Siemens Supervisory Control and Data Acquisition Systems. These systems are used to monitor and control critical infrastructure like power grids, reservoirs and nuclear power stations. The majority of Stuxnet infections were discovered in Iran, targeting nuclear centrifuges (Karnouskos, 2011).

Shortly, after the discovery of Stuxnet, a new sophisticated piece of malware named Duqu was discovered by security researchers in Hungary (Bencsáth *et al.*, 2012). This malware shared programmatic similarities to Stuxnet, but was designed for a different purpose, that is to capture passwords, documents and end user behaviour on a computer system and transmit these back to the adversary. Duqu was traced back to targets in Iran and Sudan. In both cases, security researchers discovered that these two pieces of malware used previously undiscovered vulnerabilities in the Microsoft Windows operating system to exploit targeted computer systems (Bencsáth *et al.*, 2012).

That said, two years later a more advanced version of Duqu was discovered by the security firm Kaspersky Lab, named Duqu 2.0 (Kaspersky Lab, 2015b). However, this time the target was different, being the security firm Kaspersky itself. Kaspersky, a security service provider and malware research firm that discovered the Stuxnet worm, became the target of sophisticated malware sharing the same similarities in design as its predecessors. Kaspersky reported that Duqu 2.0 also utilized previously undiscovered vulnerabilities in the Microsoft operating system to exploit computer systems on its internal network (Kaspersky Lab, 2015b). With reference to this, in a recent news article,[2] use of the Duqu 2.0 malware in the Kaspersky network breach was attributed to Israeli intelligence (Perlroth & Shane, 2017).

2.1.4 Malware attribution

Although malware can easily be attributed to a particular malware family and/or category, it is important to note that the attribution of malware to a particular criminal group and/or entity is a difficult task (Bartholomew & Guerrero-Saade, 2016). For example, malware authors could purposefully add programmatic elements to malicious software code that mimics code similarities of different cyber criminal groups, thereby implicating the wrong cyber criminal group in an attack (Bartholomew & Guerrero-Saade, 2016).

[2] https://mobile.nytimes.com/2017/10/10/technology/kaspersky-lab-israel-russia-hacking.html

2.2 Phishing

"A Good Phishing Attack is Worth a Million Zero-Days" –Thomas Fischer

What started as a systematic plan by defiant juveniles to steal passwords by email, advanced into one of the top cyber threats affecting individuals, organizations and governments today (Rekouche, 2011). Email forms a vital part of business and personal day-to-day digital communication with billions of emails sent daily over the Internet. However, this vital digital communication platform also introduced an ever changing threat to businesses and social users. Owing to its ease of use and the lack of security awareness amongst its users, this digital communication platform is also leveraged by cyber criminals for malicious operations (Meidam, 2015).

Phishing is the process whereby a cyber criminal crafts a special email message to entice an individual to disclose sensitive information that would not normally be disclosed. For example, a cyber criminal could spoof an email from a bank to an individual to obtain personal bank account information (Elledge, 2004).

Over time phishing has evolved into different types of phishing attacks tailored to specific cyber criminal objectives. For example, targeted phishing is a more personalized phishing attack whereby cyber criminals target a specific user or department in an organization. In this case, cyber criminals prepare an email attachment containing malicious software to compromise their target (Meidam, 2015). High volumes of malware driven email, known as malicious spam, is another type of phishing attack seen almost daily in organizations. Often, the malicious software employed in these types of attacks can be used to capture sensitive user or corporate data (Symantec, 2016b). In addition, an attacker could also leverage such malicious software to establish a covert channel out of a corporate network, allowing cyber criminals direct internal access to corporate assets (Elledge, 2004).

In the following subsections we discuss the financial impact of malware used in targeted phishing and malicious spam attacks, defending against phishing and related research in this field.

2.2.1 Impact of targeted phishing

One could easily argue that one of the biggest threats to an organization is targeted phishing. Cyber criminal syndicates carefully select a target organization with a particular goal in mind (Villeneuve, 2011). In recent targeted phishing attacks against global banks we have seen cyber criminals target the Society for Worldwide Interbank Financial Telecommunication (SWIFT) system (Nettitude, 2016). In February 2016, the central bank of Bangladesh confirmed that their SWIFT systems were compromised through a targeted phishing attack (Nettitude, 2016).

It was reported that in this attack cyber criminals crafted a malicious curriculum vitae document that was used to target the Human Resource department to establish an initial foothold in the network (Cheng, 2017). Cyber criminals then attempted to transfer $1 billion. However, they only managed to transfer $81 million. In December 2015 an attack following the same modus operandi was reported from a Vietnamese bank. However, the attackers did not manage to transfer any funds. In January 2015 another similar SWIFT attack was reported by the bank of Ecuador. This time, cyber criminals managed to transfer $12 million (Nettitude, 2016). In June 2016 an unnamed Ukrainian bank also reported that they had been compromised and cyber criminals managed to transfer $10 million through their SWIFT system (Nettitude, 2016).

Between August 2015 and February 2016 a cyber criminal group known as “Buhtrap” were responsible for targeting Russian and Ukrainian banking clients with malicious spam designed to compromise sensitive client banking information (Group IB, 2016). Later, moving away from malicious spam, the same cyber criminal group targeted specific Russian banks using the same malicious software seen in the malicious spam attacks. It was reported that the “Buhtrap” group managed to transfer a total of $27.5 million during these targeted phishing campaigns. Thus, as demonstrated by the “Buhtrap” group, malicious software used in mass malware attacks could also be leveraged in more targeted phishing (Group IB, 2016).

In 2014, JP Morgan, one of the largest US investment banks, reported one of the most significant data breaches resulting from a staff member having opened a targeted phishing email containing a malicious attachment (Vanier, 2016). Furthermore, the cyber criminals managed to obtain a foothold in the internal network of the bank. However, instead of compromising financial systems to transfer funds, the cyber criminals ex-filtrated the records of 83 million clients over a period of 7 months. The compromised data were used to set up an elegant fraud scheme, which lead to losses amounting to $100 million before it was terminated (Vanier, 2016).

In late 2013, "The Great Bank Robbery" was discovered by security firm Kaspersky (Kaspersky Lab, 2015a). It was reported that 100 banks globally had been targeted by a sophisticated cyber criminal group referred to as *Carbanak*. The criminal group used a targeted phishing attack to deploy malicious software known as *Carbanak* to provide them backdoor access to the targeted banks globally. In addition, the malware was designed to allow the criminal group to perform video surveillance on their targets by using the built-in video capability of the systems targeted. Through this surveillance, the criminal group could ascertain intricate business knowledge of internal banking operations, enabling them to gain access to the targeted banks' automatic teller machine (ATM) networks and to dispense cash at will. It is estimated that the criminal group stole up to $1 billion during their campaign targeting banks globally (Kaspersky Lab, 2015a). In a recent report by Symantec, security researchers discovered common attributes between the malicious software used by the *Carbanak* group and some of the attacks on SWIFT systems discussed earlier (Symantec Security Response, 2016).

2.2.2 Impact of malicious spam

Over the last 10 years, security companies have reported a downward trend in spam mail (Symantec, 2015; Robinson, 2015; Kaspersky Lab, 2016). However, spam has been replaced by a more dangerous attack vector known as malicious spam (Symantec, 2015; Robinson, 2015; Kaspersky Lab, 2016). Although malicious spam is not a novel attack vector, it allows an attacker to craft a malicious email attachment in the form of a legitimate business file that can bypass traditional anti-malware controls and ultimately compromise a user's system (Robinson, 2015).

Stone-Gross *et al.* (2011) discuss an underground economy related to a malicious managed email distribution network known as the *Cutwail* botnet. The cyber criminals managing this network provide the required infrastructure to orchestrate automated malicious spam attacks against organizations. In addition, these malicious spam services can be acquired on underground cyber crime forums by willing buyers. Therefore, cyber criminals lacking the skill level to perform targeted phishing attacks themselves can make use of such services to orchestrate attacks against organizations. Given the availability of such services it is obvious that malicious spam is a rapidly growing threat to organizations.

Currently, one of the largest active malicious managed email distribution networks on the Internet is known as the *Necurs* botnet (Baird *et al.*, 2017). This particular botnet is well known for distributing high volumes of malicious spam. Two particular malware families associated with this botnet are known as *Locky* and *Dridex*. The former malware family is a malicious piece of software known as ransomware (Baird *et al.*, 2017). However, *Locky* is only one particular strain of the malware part of the ransomware category. Ransomware has recently been wreaking havoc across different industries globally (Symantec, 2016a). The purpose of a ransomware email driven attack is to compromise a system of an organization and render specific data on the system unusable until a spesific amount of money has been paid as a ransom to the cyber criminals targeting the organization (Symantec, 2016a).

In a press release by the Federal Bureau of Investigation (FBI), it was reported that between 2014 and 2015, US companies suffered combined financial losses of up to $47 million based on malicious ransomware spam attacks (Anderson, 2016). At the time of writing, figures for direct financial losses suffered by companies in South Africa were unfortunately not available. One reason for this may be that South African companies are reluctant to report these types of attacks. This makes it very difficult to get a comprehensive view on malicious ransomware spam attacks in South Africa (Alfreds, 2016a). However, based on reports in the media, companies and users suffered undisclosed financial losses due to ransomware email attacks (Alfreds, 2016b). Furthermore, the researcher is aware of 13 confirmed ransomware incidents in his own organization during 2016. However, no direct financial losses were reported based on these 13 ransomware incidents.

It comes as no surprise that cyber criminals have gone one step further to build a business model around Ransomware-as-a-Service (RaaS) (McAfee Labs, 2016). While our discussion on the *Cutwail* and *Necurs* botnets focused on the infrastructure available in the underground to orchestrate malicious spam attacks on organizations, the RaaS business model provides an end-to-end ransomware crime service. For example, any person connected to the Internet can leverage such a service to orchestrate an attack on any organization of choice with little or no skill. The cyber criminal syndicates running these operations perform all the advanced functions in the background, whilst claiming a percentage of the ransom money acquired from the compromised victim (McAfee Labs, 2016). Therefore, one can see how any organization can become a target today and the serious threat these managed crime services pose to an organization.

In the 2014 “State of Financial Trojans report” by Wueest (2015), it was reported that almost every type of financial institution had been targeted by financial malware like the *Dridex* malware used in malicious spam attacks. In addition, cyber criminals create malicious document files embedded with the malware and mask them as legitimate documents before being distributed to the target organization.

However, not only is this malware used in high volume malicious spam attacks, but cyber criminals have also used it in targeted phishing attacks. For example, in 2015 a Swiss based organization suffered such an attack during which cyber criminals managed to transfer more than $1 million from a high value corporate account (Wueest, 2015). It was further reported that in 2015 financial organizations in the UK suffered $30 million in losses due to the *Dridex* financial malware used in malicious spam attacks. It is estimated that financial organizations across the globe suffered up to $100 million in losses due to the *Dridex* financial malware (Kharpal, 2015).

It is important to note that, although malicious spam attacks are seen as a nuisance for cyber defenders to deal with (McAfee Labs, 2016), once such an attack is successful, the information of the organization where the malware is active can be sold on underground markets. Advanced cyber criminals can buy this information on these underground markets to structure more advanced attacks on the particular organization (Wueest, 2016). For example, traditional cyber criminals would be interested in obtaining credentials to compromise bank accounts. However, advanced cyber criminals are likely to compromise additional hosts in the target network to obtain sensitive trade information or compromise financial processing systems (Wueest, 2016).

The security firm Symantec reported that similar methods were used in attacks on banks in 2016 (Symantec Security Response, 2016). Moreover, sophisticated malware was used to compromise bank machines that were initially compromised by malicious spam attacks (Symantec Security Response, 2016). In a report by Group IB & Fox-IT (2014), it was reported that in 2014 an advanced cyber criminal group known as “Anunak” purchased high value target information from cyber criminals performing malicious spam attacks on organizations. It was further stated that this particular advanced cyber criminal group transferred more than $25 million from Eastern European financial organizations in 2014.

2.3 Defence against Phishing

“You could spend a fortune purchasing technology and services, and your network infrastructure could still remain vulnerable to old-fashioned manipulation.” –Kevin Mitnick

2.3.1 Advanced malware protection systems

Malware driven email attacks have evolved at such a rate that traditional detection methods have struggled to keep up (Frost & Sullivan, 2016). For example, as discussed in the RaaS model in Section 2.2.2, cyber criminals automate the creation of new malware strains to bypass traditional malware detection systems with ease (McAfee Labs, 2016). Therefore, an organization should look at adopting an Advanced Malware Protection System (AMPS), which can provide an organization with behaviour based analysis of suspicious files entering the organization. Instead of just traditionally looking at static values of suspicious files to generate a verdict, the AMPS can execute a file in an isolated environment to monitor the programmatic behaviour of such a suspicious file (Frost & Sullivan, 2016).

It is worth mentioning that in the *Carbanak* attack discussed in Section 2.2.1, it was reported that the AMPS from Trend Micro could identify malicious elements within the particular malware used during the attack by dynamically analysing the programmatic behaviour of the malicious file (Irinco, 2015). Furthermore, in a recent report by Kaspersky Lab the security research team were notified of two separate sophisticated malware based attacks on banks based in Poland and South East Asia through their AMPS (Kaspersky Lab, 2017a). An AMPS can further protect an organization against malicious spam attacks like ransomware (Trend Micro, 2017).

2.3.2 Phishing awareness

An organization can have a multitude of technical defensive layers to combat cyber attacks (Shaw *et al.*, 2009). However, an unsuspecting employee opening a malicious attachment could result in financial losses of millions to an organization as was the case in the SWIFT (Nettitude, 2016) security breaches and *Carbanak* (Kaspersky Lab, 2015a).

In a study by Dhamija & Hearst (2006) the researchers found that a high quality phishing attack could still fool the majority of test subjects. Therefore, it is imperative that employees are educated in cyber security awareness to form part of the human cyber defensive layer in an organization (Shaw *et al.*, 2009). Employees that are trained and aware of cyber threats understand their role in protecting an organization's information. The awareness of cyber threats in an organization should promote a cyber cautious culture, whereby employees would think before opening a suspicious attachment or web link (Shaw *et al.*, 2009).

2.4 Log Data Analysis

"The ability to take data, to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it, that's going to be a hugely important skill in the next decades." –Hal Varian

Operating systems, network devices and software applications typically provide some functionality to generate computerized data messages known as "logs" when certain events occur in a system, device or application. Logs or log files are often used to identify system resource or application performance issues and provide a sequential track record of events that have taken place on a system (PCI Security Standards Council LLC, 2016).

Networking, system, security and technical logs are critical when performing cyber security monitoring and/or doing cyber security investigations. These logs provide a layer of valuable data that forms part of such monitoring and investigative activities (Crest, 2015). For example, a firewall provides information of communication between a source and destination IP. In addition, an application log could provide information about the user account used in conjunction with the source IP. Therefore, fusion of these different layers of log data can provide an intelligent log data layer that can be used in such monitoring and/or investigative activities (Crest, 2015).

Another example of valuable logs that could be used for monitoring and/or investigative activities originates from PowerShell (Symantec, 2016c). PowerShell is an interactive command line and scripting framework that forms part of modern Microsoft Windows systems used for system administration and automation. However, this framework has also been used as a tool of choice for cyber criminals to orchestrate attacks on organizations. PowerShell code is often used by cyber criminals in document files to download malware and infect a system. However, such activity can easily be detected by monitoring PowerShell logs (Symantec, 2016c).

2.4.1 Fused log data as a source of intelligent data

The fusion of multiple data sets is a popular technique in big data analytics to derive intelligent data in today's financial markets (Chen & Storey, 2012). Moreover, deep analysis of complex data sets has also been used to determine the impact of business decisions.

Fayyad *et al.* (1996) applied various data analysis techniques to discover intelligent data in business areas like finance, marketing, manufacturing and fraud. For example, the intelligent data discovered in the fraud business area could be directly applied in the detection of money laundering in the US Treasury Financial Crimes Enforcement Network (Fayyad *et al.*, 1996). Intelligent data can therefore be defined as data that can be used in the decision making process to determine the impact of a business decision (Chen & Storey, 2012).

In a survey by Goebel & Gruenwald (1999) a multitude of data analysis techniques are discussed. The authors also explain that the application of different data analysis techniques to data sets can produce different levels of intelligent information for an organization. Intelligent data can also be used in an organization to understand and mitigate the cyber threats to the organization (Hutchins *et al.*, 2011). Therefore, intelligent data can provide valuable information to an organization about its modern day adversaries and how to defend against these (Hutchins *et al.*, 2011).

2.4.2 Fused log data as a source of threat intelligence

One of the key pillars of threat intelligence is data fusion and analysis (Informatica, 2013). Data fusion can be defined as the combination of multiple data log sources to create a new threat intelligence data set (Informatica, 2013). Hawthorne (2016) discussed the importance of fusing segregated cyber data sets to provide a comprehensive view of the cyber threats targeting an organization.

With reference to this, Hoppe *et al.* (2009) discovered hidden patterns between the propagation of malware and internal computer systems. The researchers fused anti-malware solution log data with the data of internal computer system attributes to produce a new threat intelligent data set that formed part of the data analysis that uncovered a previously unseen data pattern.

Tarala (2011) discussed the importance of fusing log data as part of continuous cyber security monitoring in an organization, thereby providing cyber situational awareness of threats targeting the organization. In addition, cyber security analysts are provided with greater context within which to analyse such threats, increasing the overall efficiency of the analysts.

Oprea *et al.* (2015) fused the log data of domain name system (DNS), web proxies and firewalls to produce a threat intelligence layer to aid in the detection of malicious network traffic during the early stages of a malware infection within a corporate network.

Abraham & Nair (2015) proposed a predictive cyber security framework that uses a data analysis technique known as an attack graph. The framework combined the data of vulnerabilities usually found within network services on single network hosts with vulnerability security attributes to identify the shortest attack path to the most vulnerable information system on the network.

Arnao *et al.* (2015) developed the “Laika BOSS” intrusion detection and malware analysis platform. This platform can collect data from multiple different data log sources and fuse the data sets together with additional meta-data about potential threat actors. In addition, this platform allows security analysts to respond to new threats in a timely manner based on the different levels of analysis routines that can be performed on disparate data sets. As a result, data analysed through the “Laika BOSS” platform, can provide a security analyst with a threat intelligence layer that reduces the manual analysis time and increases the overall efficiency of the analyst (Arnao *et al.*, 2015).

2.5 Active Directory

"My message for companies that think they haven't been attacked is: "You're not looking hard enough" –James Snook

2.5.1 Overview of Active Directory

In 2000, Microsoft released Windows Server 2000 introducing a new component known as Active Directory (Microsoft, 2000). Ten years later, it was reported that AD was being used in 95% of Fortune 1000 companies (Gohstand, 2010). AD provides an organization with an active database to store, authenticate and administer organizational resources on a network. Such resources, known as *objects*, contain different data fields to identify each *object* uniquely (Microsoft, 2000). For example, a user *object* can contain business context information, like the department, job title or office location where the user resides. In addition, *objects* also contain permissions that allow the user *object* to access network resources across the network (Microsoft, 2000).

2.5.2 Active Directory in a security context

The context of cyber security related information has become an important element in cyber security monitoring and incident investigations (Crest, 2015). "There is a reason LDAP monitoring is being used by some of the most successful information security teams. It is a powerful addition to any information security program that is able to harness its true potential and resolve complex attack scenarios" (Gumbs, 2017).

As adversaries attempt to obtain an entry point into an organization by using techniques like targeted phishing to compromise high value users in an organization (Splunk, 2017), the monitoring of high value *objects* in AD has become critical for cyber security monitoring teams. For example, the domain admin *object* in AD contains privileged user *objects* with full control over an organization's internal network resources. Therefore, one can see the security attributes of these defined privileged user *objects* and the security context that AD provides (Splunk, 2017).

Furthermore, the security context could also highlight the significance of an attack should it have been successful. For example, if an adversary targeted a high value user account or system and the attack was prevented, the security context around this could highlight the potential business impact if such a user account or system had been breached (Hewlett Packard Enterprise, 2016).

2.6 Threat Intelligence

"If you know the enemy and know yourself you need not fear the results of a hundred battles." –Sun Tzu

Gartner defines threat intelligence as "evidence-based knowledge, including context, mechanisms, indicators, implications and actionable advice, about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject's response to that menace or hazard" (Webroot, 2013). However, threat intelligence is often associated with static data points provided by vendors, for example, file reputation values or command and control server information of threat actors collected by the vendors. These static data points also provide a level of actionable threat intelligence collected from cyber attacks observed across different organizations (Securosis, 2015).

2.6.1 Use of threat intelligence

Although static data points provide specific information about a particular threat, this does not necessarily provide intelligence. Threat information should be used in context and correlated with the operational environment in which it is assessed. In addition, intelligence should provide answers to drive business decisions in an organization (Walker, 2016). In the 2014 "SANS Intelligence and Analytics" survey (Shackleford, 2014), it was reported that the adoption and use of threat intelligence was still slow compared with the 2013 survey. However, it was also reported that those organizations that had implemented threat intelligence into their security program had benefited from the visibility it provided of known threat indicators.

In the 2015 “SANS Who’s using Cyber Threat Intelligence and How” survey (Shackleford, 2015), it was reported that more organizations had adopted threat intelligence to the extent of having a dedicated team looking after this function in the cyber security team. However, these teams still relied on the threat intelligence data feeds provided by vendors and security communities.

In the 2016 “SANS The state of Cyber Threat Intelligence” survey (Shackleford, 2016), it was reported that organizations had implemented threat intelligence more holistically into their cyber security program. In addition, the value of threat intelligence in a cyber security program could be quantified as it matured.

In the 2017 “SANS Cyber Threat Intelligence” survey (Shackleford, 2017), a radical shift was revealed in that almost half the organizations had a dedicated threat intelligence team. Furthermore, another shift from the consumption of threat intelligence to the production of threat intelligence among internal threat intelligence teams was reported. For example, threat intelligence teams would utilize internal systems of interest to analyse and produce threat intelligence reports, thereby not only relying on vendor feeds, but adopting a hybrid model of producing and consuming threat intelligence (Shackleford, 2017).

Although we have seen a shift in the adoption of threat intelligence in organizations over the last four years, it is important to understand the value of the threat intelligence cycle (Chismon & Ruks, 2015). In the following subsections, the threat intelligence cycle as well as four important types of threat intelligence namely, technical, tactical, operational and strategic threat intelligence (Chismon & Ruks, 2015) are discussed.

2.6.2 Threat intelligence cycle

Although threat intelligence can be used to reduce cyber risk and aid threat hunters in an organization (Walker, 2016), “Doing threat intelligence is important – but doing it right is critical” (Chismon & Ruks, 2015, p.4). The five important phases of the threat intelligence cycle are the following: requirements, collection, analysis, production and evaluation. The threat intelligence cycle evaluates the quality of the data input into the first three phases and allows the input of each of these phases to be altered until it passes the evaluation phases to produce a quality data set that can be used in an organization (Chismon & Ruks, 2015). Below we briefly discuss the five phases of the threat intelligence cycle.

Phase 1 – Requirements

Requirements are imperative to determine "what" needs to be achieved by using threat intelligence and the value this will provide.

Phase 2 – Collection

The correct data points need to be identified to collect the relevant information to achieve the requirements from phase 1.

Phase 3 – Analysis

During the analysis phase, the data from phase 2 are processed and analysed. In addition, during this phase the processed data could also be fused with different data sets to produce an intelligent data set.

Phase 4 – Production

In this phase, the processed data from phase 3 are in an acceptable state to be presented to senior management or utilized in the operational environment.

Phase 5 – Evaluation

In the last phase, the output of phase 4 is evaluated against the requirements of phase 1. If all requirements have been met, deeper analysis can commence on the data set from phase 4. If the requirements have not been met, the process returns to phase 1 to establish whether the requirements set out were realistic and/or mistakes were made during phase 2 or 3.

2.6.3 Sub areas of threat intelligence

Technical threat intelligence

Technical threat intelligence is usually associated with static data points provided by threat feeds. Moreover, this type of threat intelligence provides cyber defenders short time span indicators of compromise, such as malicious IP addresses, domains and/or hashes (Chismon & Ruks, 2015). For example, one popular community driven threat intelligence feed is the Alien Vault Open Threat Exchange (OTX)[3] feed.

Tactical Threat Intelligence

Tactical threat intelligence is usually derived from industry research around advanced persistent threats (APTs). The tools, techniques, and procedures (TTPs) used by APTs are used to enhance security monitoring. TTPs can also be used to ensure that incident response teams are aware and prepared to respond to the latest cyber threats. For example, the popular tool *Mimikatz*[4] is used to dump credentials and can be used in conjunction with legitimate tools like *PStools*[5] in an organization (Chismon & Ruks, 2015).

Operational Threat Intelligence

Operational threat intelligence provides a level of situational awareness to management of cyber attacks that the organization has experienced (Chismon & Ruks, 2015). This level of threat intelligence can be used by upper management to drive the improvement of cyber security controls, for example, understanding cyber threats better and strengthening cyber preventative controls (Rattray, 2014).

Strategic Threat Intelligence

Strategic threat intelligence provides high level information to executives and Board members. Examples of the information include the financial impact of cyber

[3] https://otx.alienvault.com/

[4] https://github.com/gentilkiwi/mimikatz

[5] https://technet.microsoft.com/en-us/sysinternals/pstools.aspx

attacks or use of a vendor supply chain that has suffered numerous data breaches (Chismon & Ruks, 2015).

2.7 Related Research

Over the years there have been a variety of studies looking at some form of email data in one way or another. One of the most popular research topics in the late 90s was email "SPAM". Although email spam was not classified as malicious, it was seen more as a nuisance and responsible for high loads of unsolicited mail traffic across the globe. Therefore, at that time a fair amount of research focused on the prevention and/or reduction of overall spam mail (Goodman *et al.*, 2007).

A plethora of research has also been done on the detection of phishing (Khonji *et al.*, 2013). The author did an extensive literature survey on phishing detection techniques and found that machine learning yielded the best results at the time. However, they also highlighted the importance of educating users in the detection of phishing email as well as the use of technology in the detection and prevention thereof.

Today the term phishing goes hand in hand with cyber crime and still plagues organizations globally. Much research has been done in this particular field of email data. In 2006, Dhamija & Hearst (2006) presented the first empirical research on why phishing attacks are so successful focusing on the lack of phishing awareness of users. In their study, the researchers fused data points like age, sex, educational level and the amount of time subjects used computers.

Jagatic *et al.* (2007) conducted a social phishing experiment between groups of university students. The researchers harvested social information of the subjects to target them in a phishing attack. Subjects were selected based on the amount of quality information available online to determine how reliably a social media context could be used to succeed with a phishing attack.

They also fused data points like age, class and major subject which provided valuable information. For example, the success rate of the phishing attacks between freshman and final year students provided similar results. In addition, the phishing attack had the highest success rate with *science* majors. Ferguson

(2005) reported similar results. However, he found that junior students were highly susceptible to phishing mails with a university grade theme.

Lee & Lewis (2011) clustered disparate advanced malware data from different types of industries to identify patterns in targeted phishing attacks. They also identified specific targeted attacks against organizations in their research. However, they noted that further research would be necessary to identify commonalities across the recipients targeted, for example, if specific recipients were targeted to compromise a particular target.

Le Blond *et al.* (2014) analysed targeted malware data from a non-governmental organization (NGO) with dealings in China. The researchers enriched the data set by fusing the recipient target email address with social media information to add more context to the data and understand the job level of users targeted at the NGO. These researchers identified all four sub areas of threat intelligence, thereby showing the value of analysing targeted malware data. However, it is important to note that this study was limited to only one NGO.

The research of Graziano *et al.* (2015) showed the level of malware intelligence that can be derived by clustering a public sandbox malware data set. The researchers found that advanced malware used in targeted attacks was submitted months before making the headlines.

Various vendor based research papers (Symantec, 2015), (Phishlabs, 2016), (Kaspersky Lab, 2016), (Symantec, 2016b) and (Kaspersky Lab, 2017b) are released annually focusing on a general breakdown of malware activity seen across different organizations around the globe. In addition, certain vendor based research focuses on specific strains of malware families as seen in (Bencsáth *et al.*, 2012), (Bronk & Tikk-Ringas, 2013), (Group IB & Fox-IT, 2014), (Kaspersky Lab, 2015a), (Symantec, 2016a), (McAfee Labs, 2016), (Group IB, 2016) and (Kaspersky Lab, 2017a).

2.8 Summary

In this chapter we discussed the background concepts relevant to the research presented in the remainder of this thesis. First, we presented an overview of malware identification, its categories and families. Next, we provided background information on the financial impact of malware used in phishing attacks ranging from the advanced to the novice cyber criminal, followed by methods for defending against phishing and the importance of user education. The analysis and fusion of log data to create intelligent data was discussed as well as the usefulness of AD data in cyber security. The concept of threat intelligence was introduced including the different sub areas thereof. Finally, research directly related to this project was presented.

Chapter 3

Experimental Methodology

"Design is not just what it looks like and feels like. Design is how it works." - Steve Jobs

Our research goal is to complement the work discussed earlier by focusing our analysis on advanced malware data from one of the big five banks in South Africa. To the best of our knowledge no similar research has been done in South Africa at this time. As discussed earlier, Lee & Lewis (2011) raised the need to contextualize advanced malware data to determine if specific recipients were targeted in an advanced malware attack. This research looks at this component by fusing AD attributes with advanced malware log data, thereby providing the required context without using social media as an data enrichment source as was done by Le Blond *et al.* (2014). Le Blond *et al.* (2014) further managed to use advanced malware mail data as a source of threat intelligence. In the latest "SANS 2017 Threat Intelligence" survey (Shackleford, 2017), it was reported that more organizations are producing threat intelligence from internal systems. Therefore, our research is intended as a contribution to what has been reported in one of the latest Threat Intelligence surveys.

This chapter details the approach and methodology used in this study. In the first sections we give an overview of the experimental approach followed as well as the data analysis process. The data collection process is discussed next, followed by the data cleaning and anonymization process and finally the data fusion process.

3.1 Experimental Approach

The exploratory data analysis (EDA) philosophy relies on the visual presentation of the underlying data instead of using data analysis and/or mathematical models, to uncover hidden patterns in the data (NIST, 2017). One of the main reasons for selecting the EDA method is the nature of this method to explore data and generate a hypothesis to investigate and explore further, rather than proofing a hypothesis or evaluating the data against a particular data model (Cox & Jones, 1981).

3.2 Experimental Setup

A high-level overview of the experimental process to gain insight into the malware driven email attacks on a financial organization in South Africa is shown in Figure 3.1. Data collection starts with the collection of three data set inputs. Once collected the data sets are pre-processed, cleaned and anonymized. Finally, the three data sets are joined and the analysis is performed.

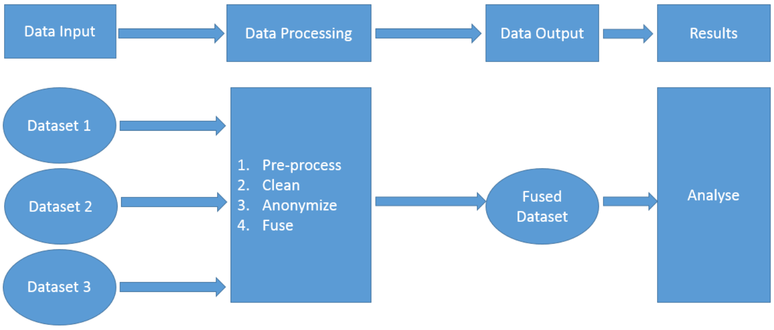


Figure 3.1: Overview of experimental approach

The primary data set used in this research has been provided by a local South African bank with the required permission for its use on condition that no personal identifiable information (PII) is published. The specific technology vendor used by the bank for the purpose of advanced email malware protection is anonymized and simply referred to as the advanced malware email protection system (AMEPS).

We fused the primary data set with malware intelligence meta-data and business context to produce the intelligent data set that was explored as a source of threat intelligence. However, to gain a deeper understanding of the individual data sets, the researcher first performed EDA on each data set individually. This was followed by the fusion of all three data sets into a new intelligent data set which was again explored using the EDA method.

The approach outlined above allowed us to explore the value that can be obtained by performing EDA on each data set individually followed by the combined value of all three data sets fused into a single data set. With the knowledge that the use of an AMEPS is not a silver bullet against the daily battle against email driven malware attacks, this research aims to contribute to a working methodology that any organization can adopt as a starting point to produce threat intelligence using advanced malware log data. As such, the research provides the first important input into the requirements phase discussed in Section 2.6.2.

3.3 Data Collection

In the following sections we explain the three data sets used in our research, including how the data were obtained and the structure of the data sets.

3.3.1 Data set 1 – AMEPS log data

The primary data set used in this research was obtained from the bank's AMEPS over a period of 6 months from December 2016 to March 2017. Our research criteria stipulated that we required all SMTP traffic logs that were flagged with malicious attachments by the bank's AMEPS during that period. The researcher was provided with a comma separated value (CSV) file export of the AMEPS data requested, containing the fields described in Table 3.1.

Table 3.1: AMEPS CSV fields

Field	Description
Time	Time stamp
Date	Date stamp
Protocol	Protocol information
Sender IP	Sender IP address
Sender	Email address of sender
Recipient	Recipient email address
Subject	Subject line used
Attachment	Attachment file name
SHA256	Cryptographic identifier of the malware identified by the AMEPS
Geo location	Geolocation from where the mail originated
Status	AMEPS verdict of the analysis of the attachment file

The query shown in Listing 1 was used to extract the information described in Table 3.1.

Listing 1 Query used to obtain email log data from the AMEPS data set

```
{"operator":"all","children":[{"field":"sample.malware","operator":"is","value":1},
{"field":"alias.email","operator":"contains","value":"@bankdomain.co.za"},
{"field":"session.tstamp","operator":"is in the range","value":["2016-12-
01T00:00:00","2017-05-31T23:59:59"]}]}
```

Our primary data set consisted of a total of 5275 malicious emails, broken down by month as shown in Table 3.2.

Table 3.2: Malicious emails detected by month

Month	Dec 2016	Jan 2017	Feb 2017	Mar 2017	Apr 2017	May 2017
Malicious emails detected	2210	583	778	976	453	725

3.3.2 Data set 2 – Malware intelligence

For the second data set, the SHA256 cryptographic identifiers associated with the malware identified in the primary data set were used. Although the SHA256 cryptographic identifier provides a unique label assigned to a particular piece of malware, the SHA256 field on its own in the primary data set does not provide sufficient information about the malware detected on the AMEPS. For this reason, a total of 2061 unique SHA256 cryptographic identifiers were extracted from the primary data set for further enhancement.

The *AutoLenz*[1] script was used to extract more contextual malware attributes from the 2061 SHA256 cryptographic identifiers. This script provides the functionality to extract different layers of meta-data about malware, for example, a user friendly name and behavioural characteristics of the malware. However, for the purpose of our malware intelligence data set we extracted only the malware attributes given in Table 3.3.

Table 3.3: Malware intelligence fields

Field	Description
SHA256	Cryptographic identifier of the malware
File Size	Reported file-size based on the cryptographic identifier
File Type	The file type associated with the cryptographic identifier
Malware Family	The malware user friendly name based on the cryptographic identifier
Malware Characteristics	Characteristics of the malware based on the cryptographic identifier

The Python script given in Listing 2 was used to extract the fields in Table 3.3.

Listing 2 AutoLenz syntax used to extract malware attribute

```
python af_lenz.py -i UniqueSHA256listfile -q "APIKey" -r meta_scrape -l 0 >
UniqueSHA256output
```

3.3.3 Data set 3 – Active Directory business context

The recipient email addresses obtained from our primary data set provided the particular target destination of the malware. However, this particular piece of information on its own does not provide the required business context. For example, an email recipient could have a specific job title and work in a particular business cluster and/or department. Alternatively, an email recipient could be a distribution list used in a particular business cluster and/or department.

[1] Available from https://github.com/PaloAltoNetworks/autofocus-lenz

Therefore, to obtain more business context around targeted mail recipients, a total of 4288 unique email recipients were extracted from the primary data set. The extracted list of email recipients was given to the bank's information security database administrator to extract the respective business context fields from AD as given in Table 3.4.

Table 3.4: Active Directory business context fields

Active Directory Field	Description
Mail	This field maps to the recipient email address targeted
Company	This field maps to the relevant business cluster in which the recipient resides
Department	This field maps to the relevant department in which the recipient resides
Title	This field maps to the relevant job title associated with the mail recipient

All recipient AD business context data were available in an SQL database managed by the bank's information security team. The query given in Listing 3 was used to extract the required data as given in Table 3.4.

Listing 3 SQL query used to extract AD business context data

```
select mail,Company,department,title from [Domain_ActiveDirectory]..Bankdomain
where mail in (select f1 from [SQL_Repository_Reporting].[dbo].[Acheckanddelete])
```

3.4 Data Pre-processing

In the real world, data sets are prone to errors and incompleteness. For example, exported data sets could have missing fields or contain invalid characters in certain fields required for analysis. Therefore, each data set was subjected to data quality checks to avoid the pitfall of garbage data in producing garbage data out. In addition, working with the private data of a bank, certain fields needed to be anonymized as per agreement with the bank. In this section we discuss the data set quality validation, cleaning and anonymization processes.

3.4.1 Data set quality validation

The primary data set was manually reviewed as it forms the foundation of this research with further data fusion based on specific fields in this data set. To start with, a timeline analysis of the exported data was performed to validate that no gaps existed in the primary data set. The timeline analysis was further validated by the time range syntax used in the query to export the primary data set. Next, the exported primary data set was checked for completeness by performing a check for any empty fields especially for the email recipient and SHA256 fields that were used in the data fusion process. Overall the primary data set was of high quality and consistent across all fields. However, the same could not be said for data sets 2 and 3, both of which required cleaning as discussed in the next sections.

3.4.2 Data set cleaning – Data set 2 – Malware intelligence

In data set 2, the researcher identified a total of 244 SHA256 values with no assigned malware classification or characteristics. The empty fields were filled in with the value "unclassified" for the purpose of this research. In addition, under the malware classification and characteristics fields, the value "Unit42" was appended to each field. This value was removed from all these fields as it served no purpose in the analysis.

3.4.3 Data set cleaning – Data set 3 – Active Directory business context

In data set 3, multiple cleaning activities were performed outlining various inconsistencies found in the data. The first activity was to identify any invalid recipient email addresses not found in the AD. Based on the data export received, the researcher was informed that all "Null" values reflected invalid email recipients. Therefore, a new column was added to this data set called "Valid Email". All email recipients with "Null" values were changed to "invalid" in this column and the remainder to "valid". This would allow the researcher to determine the validity of a targeted recipient email address more easily.

It should however, be noted that under certain circumstances a recipient address could have been valid at a particular point in time. For example, if a recipient email address was active in the first two months of the captured research data after which the staff member resigned and/or the distribution list was decommissioned, then this address could have been valid at that time. However, to determine that level of information, an additional data source such as a Human Resource Management System would be required. This is beyond the scope of this project.

Owing to the high number of email distribution lists identified in this data set, no descriptive information was available for analysis. Therefore, the researcher populated the "job title" field for all distribution lists identified with the value "Distribution list". A "Distribution list" was subsequently defined as an email recipient not conforming to the standard user naming convention provided by the bank. The same method was used for "branch manager distribution lists" although these could be easily distinguished based on the naming convention used in the recipient email address. For example, the "branch manager distribution lists" contained the suffix "BM" in the email recipient address. Therefore, the researcher populated the "job title" field for all "branch manager distribution lists" with the value "Branch Manager: DL".

Furthermore, it was found that abbreviations were used in certain job titles. For example, a recipient could have a job title containing "Snr" or "Senior" followed by the job role in the actual department. Therefore, to obtain a better level of consistency for analysis on data set 3, the changes set out in Table 3.5 were applied.

Table 3.5: Data set 3 cleaning tasks

Original Value	New Value	Reason
Jnr	Junior	Provide consistency for "Junior" level job titles
Snr	Senior	Provide consistency for "Senior" level job titles
Mngr	Manager	Provide consistency for "Manager" level job titles
Head of Department	Head	Provide consistency for "Head" level job titles
Department Executive	Executive	Provide consistency for "Executive" level job titles
C-Level Roles	Chief	Provide consistency for "Chief" level job titles

3.4.4 Data anonymization

As per the agreement between the researcher and the bank, certain data anonymization tasks were completed to anonymize the data. The changes set out in Table 3.6 were applied across data sets 1 and 3 to remove all references to the bank.

Table 3.6: Data anonymization tasks

Bank Identification Value	Replaced Value	Reason
Bank Email Domain	bankdomain.co.za	Remove all valid email bank domain references
Active Directory Domain	bankdomain	Remove internal naming of AD domain
Bank Name in Business Cluster	No Value	Removed reference to the bank
Bank User Identification	User	Remove internal user naming convention

3.5 Data Fusion

By fusing additional contextual data sets, an organization can derive intelligent information from the combined output of the different data sets (Informatica, 2013). A summary of the three data sets used in the fusion process is given in Table 3.7.

Table 3.7: Research data set information

Data set file name	Data set reference	Data set content
AMEPS	Data set 1	Six months of malicious email data identified by the AMEPS
AFlenz	Data set 2	Malware intelligence meta-data
ADdata	Data set 3	AD business context of all recipients targeted with malware

Here we discuss the process of fusing data set 1 with data sets 2 and 3. The fusion process of different data sets requires a common field and/or attribute in each data set that can be used to achieve this (Tableau, 2017b). We made use of the inner join operation to fuse the data sets. The inner join operation allowed us to fuse the different data sets based on a common field present in the data sets (Tableau, 2017b). In data set 1, the SHA256 cryptographic identifier identifies the malware associated with the email. In data set 2, the SHA256 cryptographic identifier is associated with the malware file size, file type, family and behavioural characteristics. Therefore, we used the SHA256 cryptographic identifier to perform an inner join operation on data set 1 and data set 2, thereby, enriching data set 1 with malware intelligence as shown in Figure 3.2.

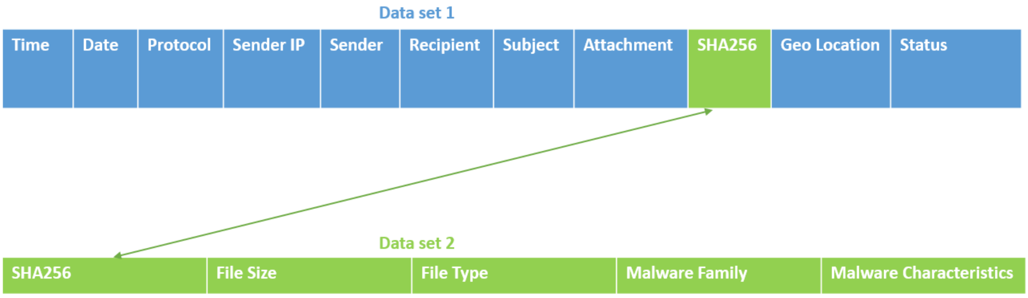


Figure 3.2: Data set 2 inner join operation

In data set 1, the email recipient field identifies to whom the malicious email was destined while in data set 3 the email recipient is associated with AD business context attributes like business cluster, department and job title. Therefore, the email recipient field was used to perform an inner join operation on data set 1 and data set 3. The inner join operation allowed the fusion of data set 1 with data set 3, thereby, enriching data set 1 with AD business context as shown in Figure 3.3.

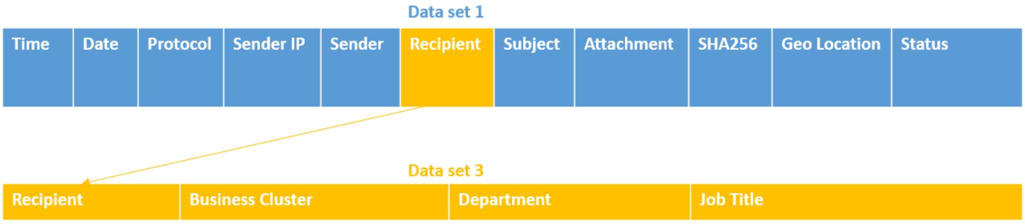


Figure 3.3: Data set 3 inner join operation

Finally, our fused data set contained two layers of context marked in green and orange as shown in Figure 3.4.

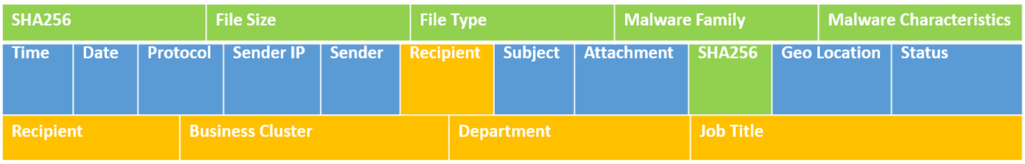


Figure 3.4: Fused data set

3.6 Hardware and Software Configuration

In this section, we provide a brief overview of the hardware and software used to conduct the research. All processing was carried out on the devices listed in Table 3.8.

Table 3.8: Data processing hardware

Device	Dell Latitude E7450	Macbook Pro 2016
Operating System	Windows 8.1	MacOS Sierra 10.12.15
CPU	Intel i7 2.7Ghz	Intel i7 2.7Ghz
Memory	8Gb	16Gb
Hard Drive	200Gb SSD	512Gb SSD

Tableau[2] version 10.3 was used as the data analysis platform. Tableau is a popular big data analysis platform used across the globe for organizations to understand their data better. Furthermore, since all three data sets used in this research were in the CSV format, the ability of Tableau to load each data set as a connection without the need to write any data to a database for analysis made it the perfect platform for data analysis (Tableau, 2017a). In addition, the researcher made use of the standard agreement between the tertiary institution and Tableau to leverage an academic license for research purposes.

3.7 Summary

This chapter described the approach followed in the collection, processing and anonymization of the three data sets used in this research. This was followed by a discussion of the data fusion process employed to create a new intelligent data set for analysis. In Chapter 4, we discuss the results of the analyses of the three data sets individually. This is followed by a discussion of the analyses of the fused data set in Chapter 5.

[2] https://www.tableau.com/

Chapter 4

Results of Individual Data Set Analyses

"In much of society, research means to investigate something you do not know or understand." -Neil Armstrong

This chapter deals with the analysis of the three individual data sets discussed in Section 3.3. EDA usually involves filtering and counting fields in a data set. These actions form part of the foundation of generating visualizations to explore and interpret the results of data analysis (Kirk, 2016). Furthermore, domain knowledge plays a vital role in the interpretation of the data. For example, a domain expert can ask a very specific question that will produce a particular result. However, for non-domain experts it is important to identify the important aspects of the data by first working through different iterations of the data. As a result, domain knowledge allow users to progress faster to more meaningful questions (Kirk, 2016).

In the following sections we discuss the results of the analysis of the three data sets. The results of each data set are discussed individually to understand the level of intelligence that can be obtained from a single data set. Sections 4.1 to 4.3 explore the results of data sets 1 to 3, respectively.

4.1 Analysis of data set 1 – AMEPS log data

The primary data set formed the foundation of the EDA. This data set contained the log information of 5275 malicious emails detected over a 6 month period between December 2016 and May 2017. In the following subsections, the analysis results of this data set are discussed.

4.1.1 Monthly malicious email analysis

To start with, a monthly analysis of the SHA256 cryptographic identifiers over the 6 month period between December 2016 and May 2017 was performed with the results illustrated in Figure 4.1. We observed a high volume of malicious emails received during December 2016 with a marked decline in January 2017. For the remainder of the period under investigation, there were slight fluctuations.

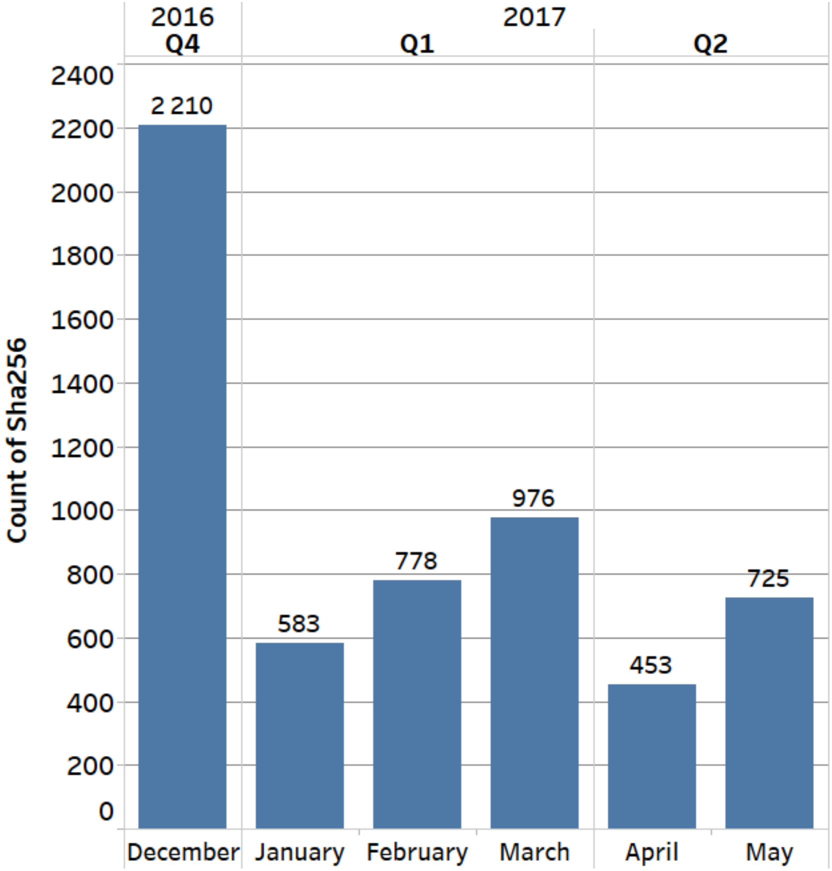


Figure 4.1: Monthly malware volume trend

In the “Spam and Phishing 2017 Q1 Report”, Kaspersky Lab reported similar findings with a sudden decrease in malicious emails globally between January 2017 and March 2017 (Kaspersky Lab, 2017b). Furthermore, in the “Internet Security Threat Report Volume 22” report, Symantec reported that the *Necurs* botnet was inactive between December 24, 2016 and March 2017 (Symantec, 2017). The *Necurs* botnet, as discussed in Section 2.2.2, is one of the largest malicious spam botnets in the world used by cyber criminals to distribute different types of malware like ransomware and financial malware.

In addition, we performed a daily trend analysis of malicious emails received over the 6 month period as shown in Figures 4.2 and 4.3. We observed a sharp decline in malicious emails on December 23, 2016 followed by a spike in malicious emails on March 29, 2017. Our results correlate with the dates associated with large changes in email volume reported by Symantec (2017).

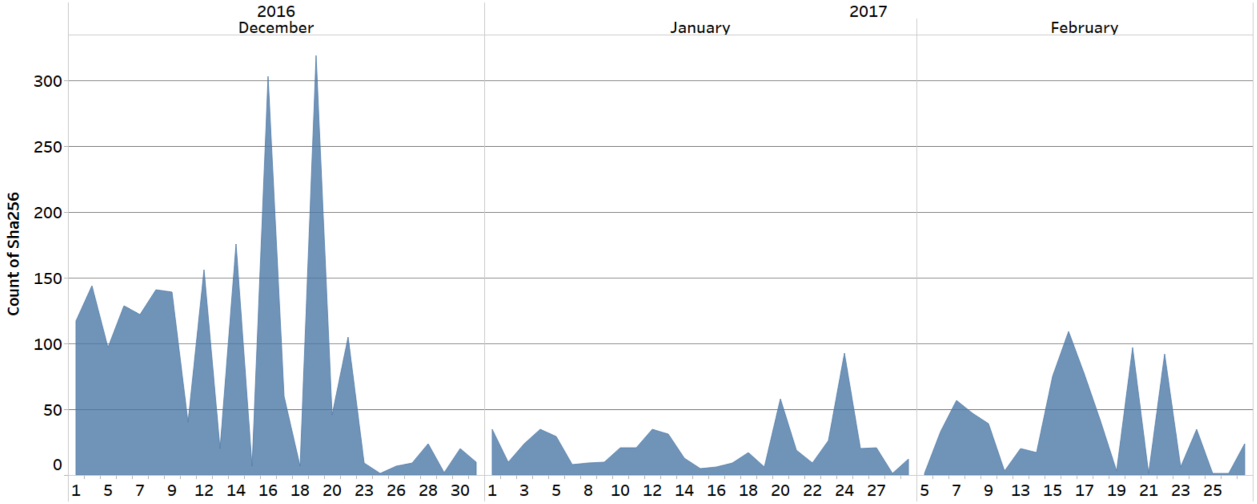


Figure 4.2: Daily malware volume Dec 2016 - Feb 2017

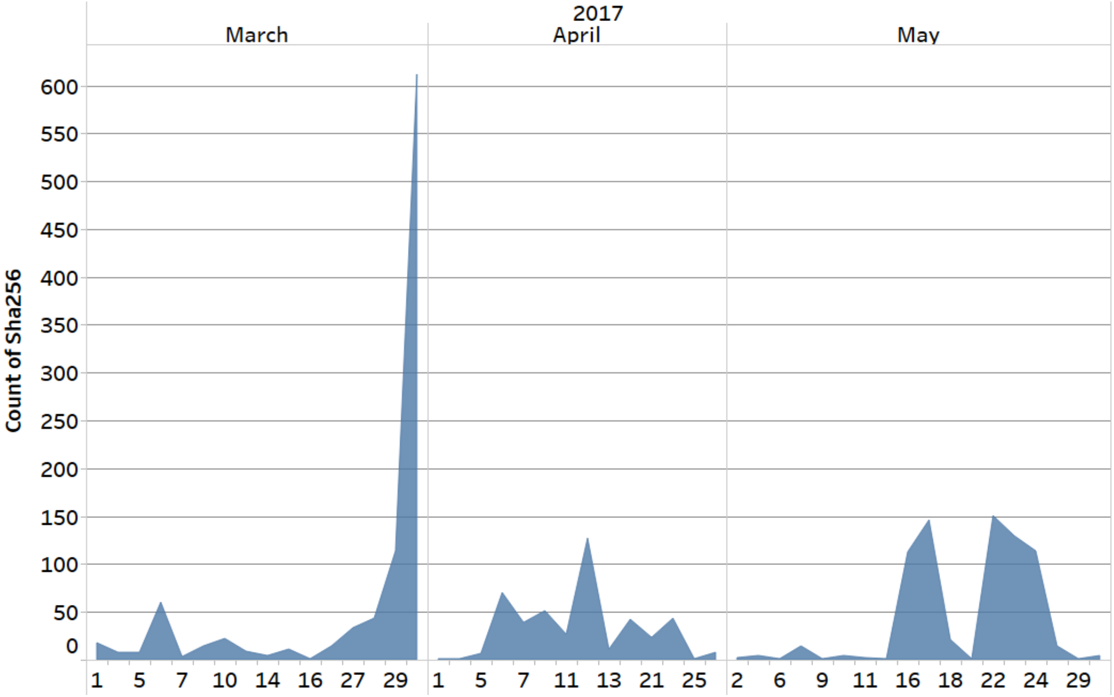


Figure 4.3: Daily malware volume Mar 2017 - May 2017

4.1.2 Situational awareness

In a recently published cyber intelligence article discussing where the busiest spammers come from (Kessem, 2017a), security researchers from IBM analysed a malicious email data set obtained between December 2016 and June 2017. The data set contained malicious email data from North America, South America, Europe and China. However, there was no mention of the African continent or South Africa in particular. The researchers looked at the day of the week, time of the day and geographic source of the distribution of malicious emails targeting these countries as situational awareness is an important aspect of threat intelligence. For example, cyber criminals will target employees during a particular time frame to ensure that the employees open the malicious emails during work hours (Kessem, 2017a). In the following sections we discuss the situational awareness findings of our data set compared with those reported by Kessem (2017a), thereby providing a financial sector perspective of situational awareness in South Africa.

Day of the week

In the cyber intelligence article by Kessem (2017a), it was reported that Tuesday, Wednesday and Thursday were the most targeted days of the week for malicious emails. However, based on our data Wednesday, Thursday and Monday were the most targeted days as shown in Figure 4.4.

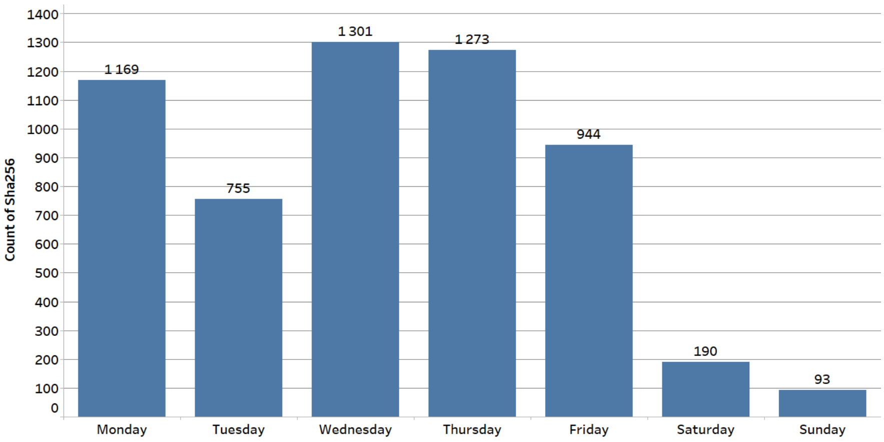


Figure 4.4: Most targeted days of the week

Time of day

Kessem (2017a) reported that malicious spam activity aligns with the normal working hours and time zones of the targeted countries. For example, one such finding was a sharp increase in malicious emails around 5am Universal Coordinated Time (UTC) with a decrease in malicious emails in the late afternoon. Our results match these in that we observed an increase in malicious emails between midnight and 5am South African local time followed by a sudden decrease between 5am and 6am. Between 6am and 8am we observed another increase in malicious email which aligns with the start of the South African business day. This was followed by a steady decrease until 11am with a slight increase over the local lunch time in South Africa at 12pm. From 12pm onwards we observed a steady decrease in malicious emails as shown in Figure 4.5.

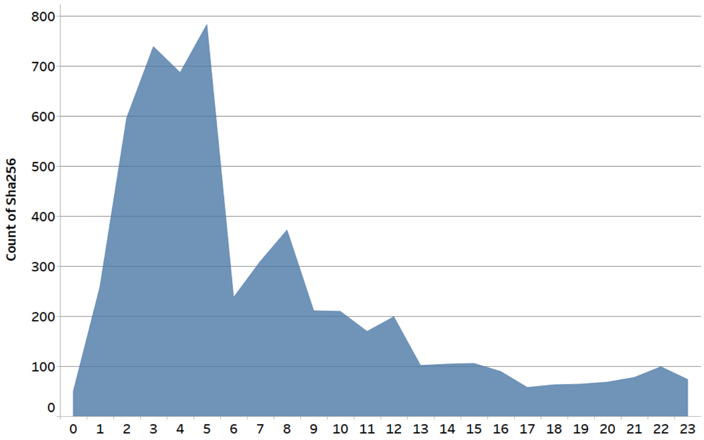


Figure 4.5: Time of day

Geographic source of malicious emails

The geographic source of malicious emails is another important aspect to consider in threat intelligence. For example, cyber criminals will typically distribute malicious emails from within the victim's own country to appear more legitimate and bypass any potential geographic mail gateway filters (Kessem, 2017a). Kessem (2017a) also reported that India, South America and China were the top three countries from where malicious spam originated between December 2016 and June 2017. Our results provided a more fine-grained view on the source countries from where malicious emails originated as shown in Figure 4.6.

According to this figure, India was the top source of malicious emails, South America was ranked second when combining the numbers of emails originating in Brazil and Argentina, followed by Russia third. Thus, our results are similar to those reported by Kessem (2017a) with the exception that Russia replaces China in third place. In addition, it is worth mentioning that the number of malicious emails originating from within South Africa was ranked seventh.

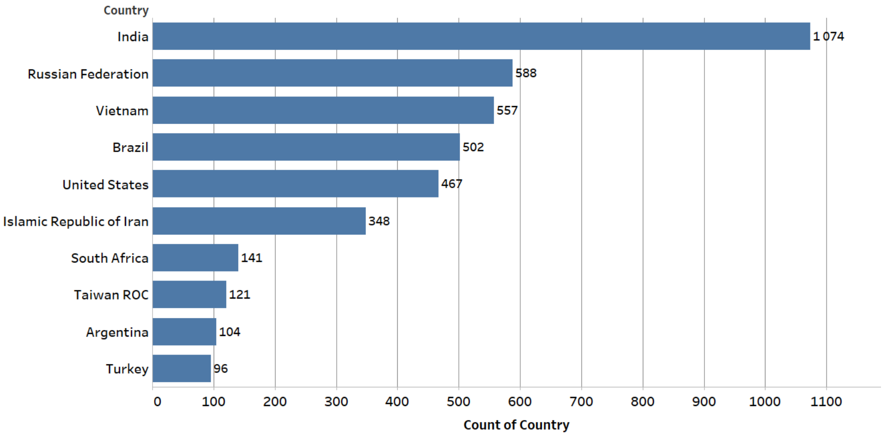


Figure 4.6: Top 10 geographic malicious email sources

4.1.3 Subject line key word analysis

The use of financial keywords has become more prevalent in malicious spam campaigns, which could reflect the potential success that cyber criminals have achieved enticing users to open malicious financial documents. In Symantec's Internet security threat report (Symantec, 2016b) it was documented that the keyword "invoice" was the top word used in malicious spam campaigns. This was followed by malicious email campaigns spoofed from an organization's internal fax machine or scanner with the keywords "document" or "scan". Finally, the third most popular keywords "mail delivery failure" were used under the pretense that a user's email had not been delivered enticing them to open the malicious mail (Symantec, 2016b).

Owing to the high number of different subject lines used even after the data cleanup, we only looked at the top 20 subject lines as shown in Figure 4.7. In our results we found that the "No Subject" line was at the top followed by the "Invoice" and "uk_confirmation" subject lines. We observed similarities in the use of keywords like "document", "Scan" and "Returned mail: see transcript for details" as reported by Symantec (2016b).

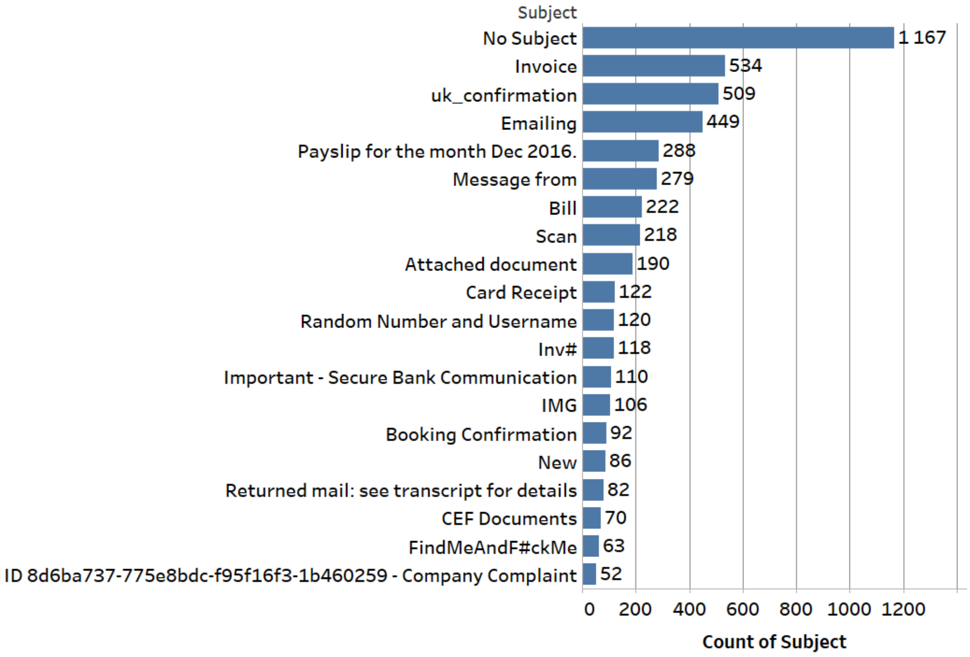
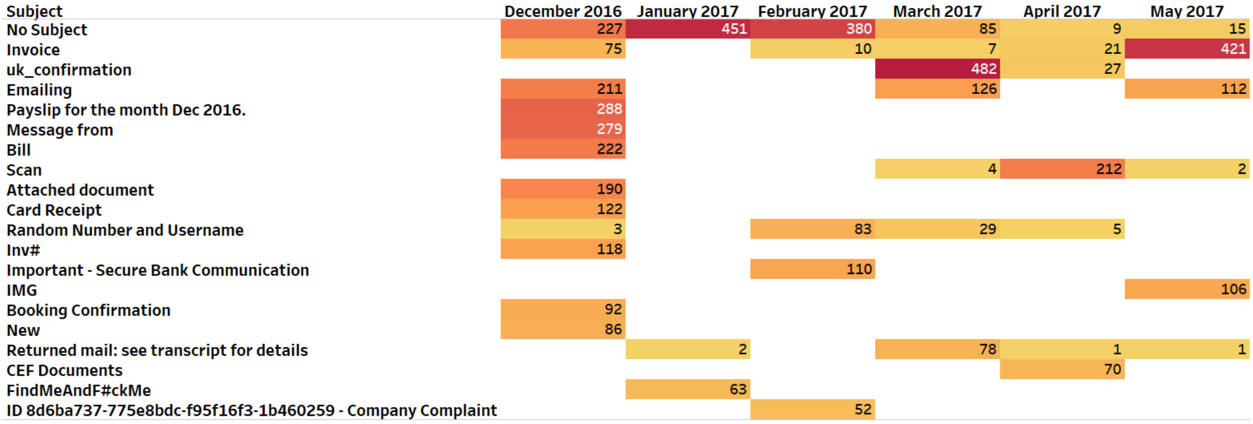


Figure 4.7: Top 20 subject lines

In addition, we looked at the distribution of the top 20 subject lines between December 2016 and May 2017, results of which are shown in Figure 4.8. The “No Subject” line was used every month to distribute malware, while the “invoice” subject line was used every month except January 2017. We also found a couple of unique subject lines used during certain months. Subject lines such as “Payslip for the month of Dec 2016”, “Message from”, “Bill”, “Card Receipt”, “Inv#”, “Booking Confirmation” and “New” were only used in December 2016. The use of the subject line “FindMeAndF#ckMe” was unique for January 2017, while “Important - Secure Bank Communication” and “ID 8d6ba737-775e8bdc-f95f16f3-1b460259 - Company Complaint” were unique to February 2017. “CEF Documents” was only found in email sent during April 2017 and likewise “IMG” was unique to May 2017. Therefore, with the exception of March 2017, we found a unique subject line each month based on the top 20 subject lines over the 6 month period.



Subject	December 2016	January 2017	February 2017	March 2017	April 2017	May 2017
No Subject	227	451	380	85	9	15
Invoice	75		10	7	21	421
uk_confirmation				482	27	
Emailing	211			126		112
Payslip for the month Dec 2016.	288					
Message from	279					
Bill	222					
Scan				4	212	2
Attached document	190					
Card Receipt	122					
Random Number and Username	3		83	29	5	
Inv#	118					
Important - Secure Bank Communication			110			
IMG						106
Booking Confirmation	92					
New	86					
Returned mail: see transcript for details		2		78	1	1
CEF Documents					70	
FindMeAndF#ckMe		63				
ID 8d6ba737-775e8bdc-f95f16f3-1b460259 - Company Complaint			52			

Figure 4.8: Distribution of the top 20 monthly subject lines

4.1.4 Summary of information from data set 1

The results of data set 1 provided a layer of situational awareness. We identified the most frequently targeted month, day and hour the bank experienced malicious email attacks as well as the top source countries from where these emails originated. Lastly, we identified the top subject lines used by cyber criminals and the uniqueness thereof.

4.2 Analysis of data set 2 – Malware intelligence

In the malware intelligence data set we extracted the meta-data of 2061 unique SHA256 cryptographic identifiers found in the 5275 malicious emails from data set 1. The meta-data allowed a particular malware family and/or exploit code to be associated with a particular cryptographic identifier, thereby providing more context of the particular malware family and/or exploit code used in the attack.

For the purpose of our research project we labeled our malware cryptographic identifiers. The first priority was to associate a particular malware cryptographic identifier with a malware family. The second priority in the absence of identifying a link between the cryptographic identifier and malware family was to use the associated behaviour identified by *AutoLenz*. Finally, if neither of these priorities could be satisfied the malware was labeled as "Unclassified". In the following sections we discuss the top ten malware families identified, followed by a discussion of the remainder of the malware families identified and broken down into seven categories.

4.2.1 Top 10 malware families identified

Half of the malware observed in the top 10 malware families belonged to the ransomware category as shown in Figure 4.9. For example, *Locky*, *CerberSage_Distribution* and *Cerber* are all well known ransomware distributed by the *Necurs* botnet (Symantec, 2017). Furthermore, in the May 2017 emerging threat report by Burbage & Kremez (2017) it was reported that the *Necurs* botnet switched to the *Jaff* ransomware family utlising a malicious *PdfDocmDropper*. Our results included both these malware families.

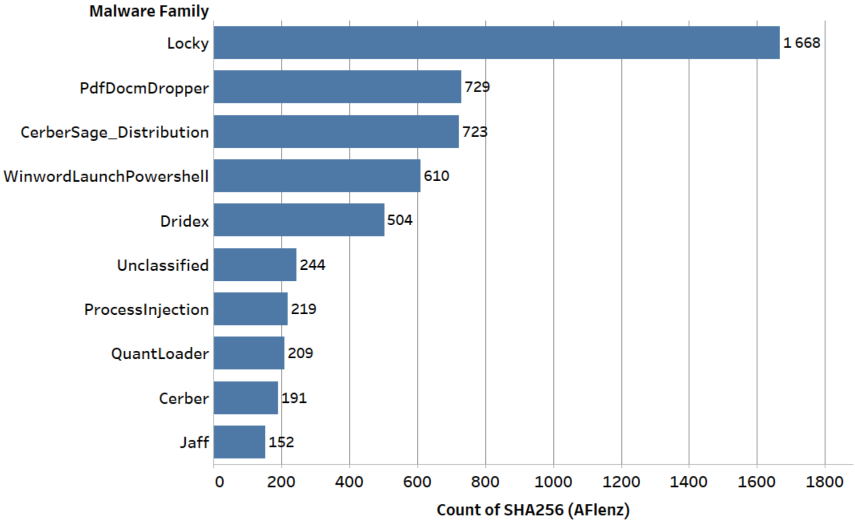


Figure 4.9: Top 10 malware families identified

The *WinWordLaunchPowerShell* and *ProcessInjection* classification did not provide us with a particular malware family. However, these behavioural attributes did yield an operational layer of threat intelligence as discussed in Section 2.6.2. For example, *WinWordLaunchPowerShell* indicated that the word document was utilising PowerShell to perform some kind of malicious action. Therefore, from an operational threat intelligence perspective an organization should ensure that they had the required PowerShell logging enabled for security monitoring purposes (Symantec, 2016c). In addition, *ProcessInjection* highlighted another valuable behavioural attribute for tactical and operational threat intelligence discussed in Section 2.6.2. For example, the behavioural attribute of *ProcessInjection* indicated that the malware could execute malicious code in the context of a legitimate system process. This technique is often used by malware to blend in with legitimate system processes, thereby evading detection (MITRE, 2017a). However, an organization could make use of a free system monitoring tool like *sysmon*[1] to detect such behaviour on a system (Russinovich, 2016).

Our results further revealed the well known *Dridex* financial malware used to steal financial information from users globally (Symantec, 2017). Just like *Locky* and *Cerber*, the *Dridex* malware has also been linked to the malicious spam distribution of the *Necurs* botnet as discussed in Section 2.2.2. We also found a number of "Unclassified" malware samples, which indicated that the researcher could not find the relevant meta-data about the cryptographic SHA256 identifiers using the *AutoLenz* script. It should also be mentioned that a couple of random "Unclassified" SHA256 cryptographic identifiers were searched for on malware intelligence websites like *Virustotal*[2], *Hybrid-Analysis*[3] and *Malwr*[4]. However, none of these widely used public malware intelligence websites had any information about the malware. Therefore, the "Unclassified" malware samples would require further analysis which is out of scope for this research.

Finally, we observed the *QuantLoader* malicious dropper, which reportedly has been sold on Russian cyber criminal markets (Griffin, 2016). This malicious dropper has also been linked to the distribution of ransomware and financial malware.

4.2.2 Remainder of the malware families

The top 10 malware families were excluded from the results of the remainder of the malware families as shown in Figure 4.10. Owing to the variety of malware families found, our discussion thereof is based on seven malware categories namely, exploit code, behavioural attributes, ransomware, information stealers, financial malware, malicious downloaders and legacy malware.

[1] https://docs.microsoft.com/en-us/sysinternals/downloads/sysmon

[2] https://www.virustotal.com

[3] https://www.hybrid-analysis.com/

[4] https://malwr.com/

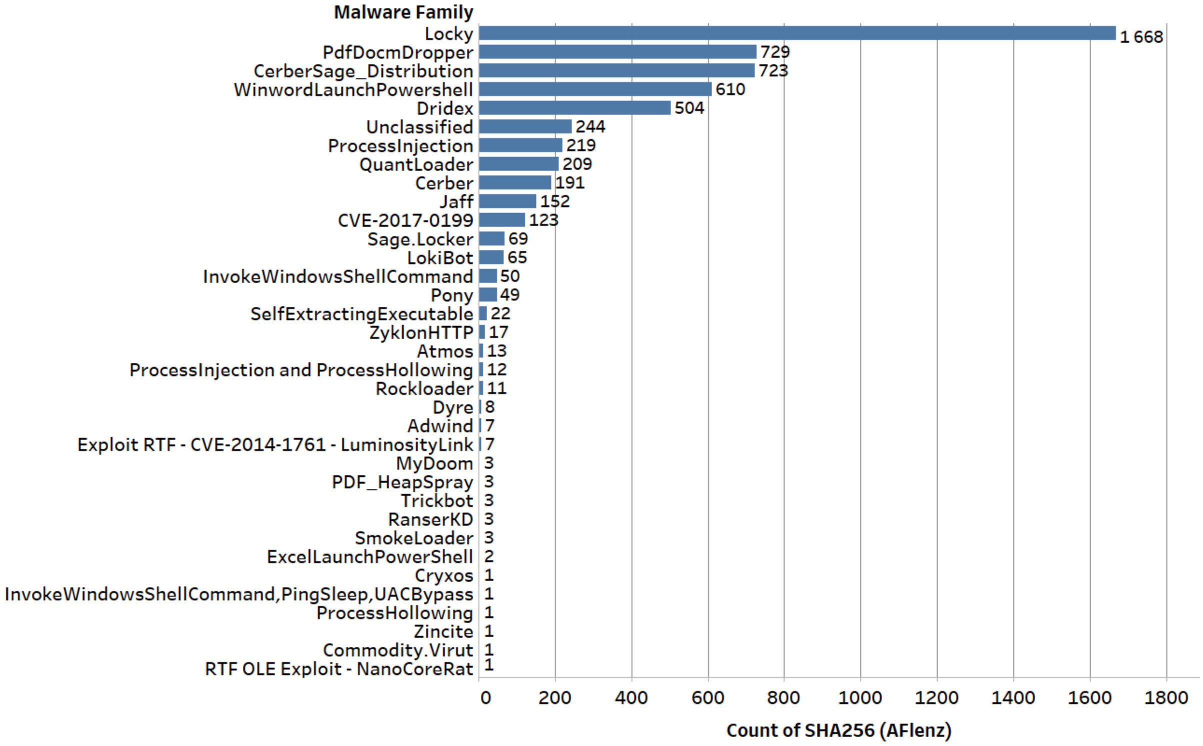


Figure 4.10: Remainder of the malware families identified

Category 1 – Exploit code

The discussion of the remaining malware family data begins with the exploit code observed. Although exploit code is not necessarily malware, it has been used over the years to exploit vulnerable applications with the goal of delivering and executing malware (Zamora, 2017). Firstly, we observed "CVE-2017-0199" at the top of the list as given in Figure 4.10. On the 11th of April 2017, Fire eye[5] reported that attackers had been exploiting a Microsoft[6] document vulnerability known as "CVE-2017-0199" (Jiang *et al.*, 2017). Furthermore, in a report by Sophos[7] it was reported that "CVE-2017-0199" had been used in targeted attacks between March 2017 and April 2017 (Szappanos, 2017). It was also found that malicious spam distribution networks were utilising this vulnerability in the distribution of financial malware like *Dridex* (Szappanos, 2017).

Secondly, "CVE-2014-1761" associated with the *LuminosityLink* malware was

[5] https://www.fireeye.com
[6] https://www.microsoft.com
[7] https://www.sophos.com

found. In a 2015, security firm Proofpoint[8] reported that the "CVE-2014-1761" vulnerability was the second most widely exploited vulnerability by cyber criminals (Bilen, 2015). Furthermore, in a 2015 research paper exposing the Cuckoo miner campaign, it was reported that Nigerian cyber criminals were targeting banks utilising a combination of software vulnerabilities like "CVE-2014-1761" and malware like *Luminosity Link* (Trend Micro, 2015).

A third family of exploit code observed was "PDF_HeapSpray". The "Heap Spray" exploitation technique is usually associated with the exploitation of Internet browsers or malicious JavaScript embedded in PDF documents (Ratanaworabhan *et al.*, 2009). However, since no information about the three SHA256 cryptographic identifiers associated with this detection could be found, it would seem to be unique and require a more in-depth analysis in future work.

Finally, we observed the "RTF OLE Exploit - *NanoCoreRat*". Although we do not have an industry identifier for the particular exploit code used, it was reported that the *NanoCoreRat* had been used in targeted attacks against the energy sector in the USA and Canada utilising older Microsoft document vulnerabilities like "CVE-2012-0158" (Llascu, 2015). However, in a more recent report, security researchers from Palo Alto Networks[9] reported *NanoCoreRat* malware activity across the Europe Middle East and Africa (EMEA) region (Hinchliffe, 2017).

The use of exploit code in email attacks provides a layer of operational threat intelligence as discussed in Section 2.6.2. Thus, organizations should ensure that systems are updated and protected from such exploit code (Zamora, 2017).

Category 2 – Behavioural attributes

In Section 4.3.1 we discussed the *WinWordLaunchPowerShell* and *ProcessInjection* behavioural attributes. However, in the remainder of the data we observed a few more behavioural attributes. For example, *ExcelLaunchPowerShell* could indicate that a malicious spreadsheet was utilising PowerShell to perform some sort of malicious action (Symantec, 2016c). Moreover, in the case of *InvokeWindowsShellCommand*, the malware could make use of PowerShell to use the "invoke" operator to execute malicious code (Symantec, 2016c).

[8] https://www.proofpoint.com

[9] https://www.paloaltonetworks.com/

A further technique used by malware known as *ProcessHollowing* to evade detection was also observed. This technique is used by malware to run a legitimate process and keep it in a suspended state whilst the malicious code executes inside the legitimate process (MITRE, 2017b). The *ExcelLaunchPowerShell*, *InvokeWindowsShellCommand* and *ProcessHollowing* attributes provide tactical and operational threat intelligence as discussed in Section 2.6.2. For example, being aware of the malware attributes discussed, an organization should have the required PowerShell logging and system monitoring in place to detect such behaviour as mentioned in Section 4.2.1.

Finally, we observed the *SelfExtractingExecutable* behavioural attribute. This particular behavioural attribute could indicate malicious code compiled into a legitimate executable file. For example, an attacker could hide malicious code in some sort of legitimate application and when the user opens the application, the malicious code automatically installs without the user's knowledge (Merrit, 2017).

Category 3 – Ransomware

When considering the lower volumes of ransomware received, we found *Sage.Locker* a known ransomware associated with the *Pandex* botnet distribution network as reported by Symantec (Altares, 2017). *RanserKD* which is another type of ransomware reported by Palo Alto Networks was also found, albeit distributed at a much smaller scale than ransomware like *Locky* and *Cerber* (Hinchliffe, 2017). Finally, we observed the *Cryxos* ransomware family (Telus Security Labs, 2016) with only one instance recorded in the data over the 6 month period.

Category 4 – Information stealers

The first observation in this category was the *ZyklonHTTP* malware, which has been distributed by malicious spam networks (Svajcer, 2017). The main functionality of *ZyklonHTTP* is to perform distributed denial of service (DDoS) attacks (Svajcer, 2017). However, it has also been reported that the malware contains the functionality to steal credentials and/or crypto currency wallets (Svajcer, 2017). *LokiBot* malware, which is another credential stealer with the functionality to steal crypto wallets similar to *ZyklonHTTP*, but without the DDoS functionality, was also found (Zhang & Liu, 2017). The *Pony* data stealing malware known for its powerful credential stealing functionality was also found (Avast, 2014). Finally, the *AdWind* malware, which has been attributed to the criminal underground Malware as a Service (MaaS) was observed (Wu & Chen, 2017). Its ease of use has made this malware family popular with novice cyber criminals, however, it has been reportedly used by advanced cyber criminal groups as well (Wu & Chen, 2017). For example, in 2016 it was reported that the *AdWind* malware was used in a targeted attack on a bank in Singapore[10] (Leyden, 2016).

Category 5 – Financial malware

In the financial malware category, *Atmos*, *Dyre* and *Trickbot* were observed. It has been reported that the *Atmos* financial malware was used in attacks against banks in France in 2016 as well as to download *Teslacrypt,* another form of ransomware (Zaharia, 2016). The *Dyre* financial malware is an older version of the *Dridex* financial malware distributed by the *Necurs* botnet (Symantec, 2017).

Finally, *Trickbot* is a new financial malware family that has been reportedly distributed by the *Necurs* botnet (Kessem, 2017b). This particular malware has been seen targeting clients of banks across the globe, albeit with no mention of South African targets (Kessem, 2017b). To verify that South African banks were actually targeted, the configuration file used by the malware would need to be reverse engineered to determine the actual targets. This is beyond the scope of this research project.

Category 6 – Malicious downloaders

In this group, the *RockLoader* and *SmokeLoader* malicious downloaders were observed. The *RockLoader* malware has been used in malicious spam campaigns to download malware like *Dridex*, *Locky* and *Pony* (Wakelin, 2016). The *SmokeLoader* malware was developed in 2011 and as recently reported, used in the distribution of financial malware like *Trickbot* (Davison, 2017).

[10] https://www.theregister.co.uk/2016/02/08/adwind/

Category 7 – Legacy malware

In the legacy malware category, we observed the *MyDoom*, *ZinCite* and *Commodity.Virut* malware. In 2004, the *MyDoom* malware was reportedly one of the most notorious mass email worms accounting for one out of twelve malicious emails at the time (Munro, 2004). The *ZinCite* malware which formed part of later *MyDoom* attacks in 2004, was classified as a backdoor, downloaded through the *MyDoom* malware, which would allow an attacker remote access to an infected machine (Leyden, 2004). Finally, the *Commodity.Virut* which was reported to be one of the most active Internet threats back in 2006[11] (Krebs, 2013), also contained a backdoor component that would allow an attacker remote access to an infected host (Symantec, 2007).

4.2.3 Summary of information from data set 2

The results of data set 2 provided a rich layer of malware intelligence based on the SHA256 cryptographic identifiers obtained from data set 1. We identified many different types of malware and exploit code used in malicious email attacks on the bank. In addition, we correlated our findings to industry examples. For example, a high concentration of malware originated from botnets. However, certain exploit code and malware were previously reported in targeted attacks around the globe.

4.3 Analysis of data set 3 – Active Directory

In simulated phishing attacks, business context plays an important role in understanding whether a particular department in an organization is improving in terms of security awareness (Greaux, 2013). For example, if a particular department in an organization remains susceptible to opening phishing emails, that area would require more security awareness training. However, the same concept could be applied to an organization's malicious email data. For example, the use of meta-data from AD like the business cluster, department and job title provides business context of "for whom" the malicious emails were destined.

[11] https://krebsonsecurity.com/2013/01/polish-takedown-targets-virut-botnet/

In data set 3 we used the 4288 unique bank email addresses from data set 1 to extract business context including the business cluster, department and job title from AD, thereby extracting business context on “who” the targets were. In the following sections we discuss the various business clusters, departments and job titles targeted.

4.3.1 Business clusters targeted by malware

As shown in Figure 4.11, the “RETAIL” business cluster was the most targeted cluster in the bank during the period December 2016 to May 2017. It is important to note that this particular business cluster contains the largest number of employees in the bank. In addition, we observed a high volume of invalid email addresses targeted. This could indicate that old bank email addresses were being used in malicious spam bot networks like the *Necurs* botnet for example.

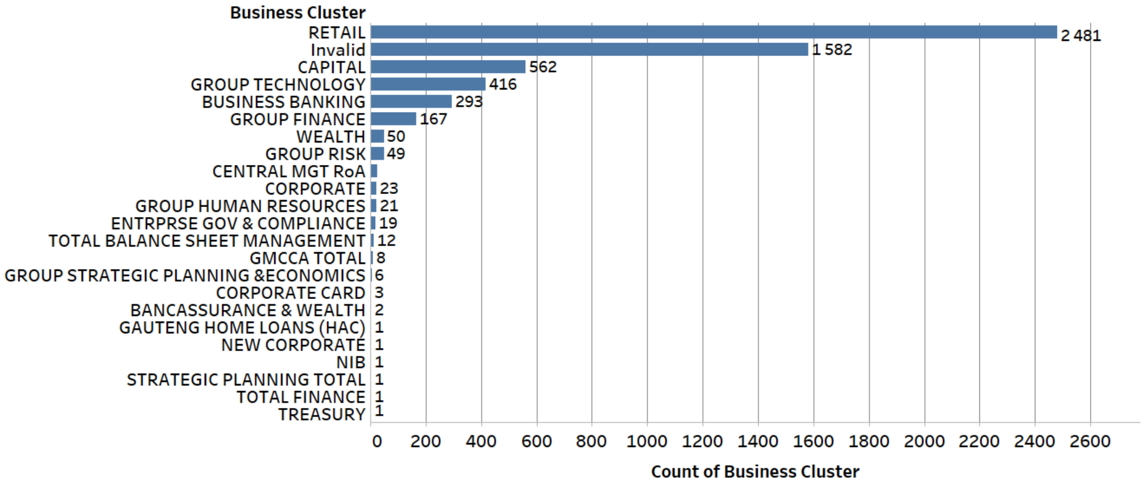


Figure 4.11: Business clusters targeted by malware

Going one step further and considering the top 20 job titles targeted within the “RETAIL” business cluster as shown in Figure 4.12, it is clear that “Distribution Lists” and “Branch Manager: DL” were the most targeted internal distribution groups between December 2016 and May 2017. Internal distribution groups provide an advantage to attackers, as security controls applied to these groups are often relaxed (Cove, 2010). In addition, targeting internal distribution groups provides the attacker a deeper reach into the organization by targeting unknown employee email addresses that form part of the internal distribution group (Cove, 2010). In a recent article by Proofpoint (2017), the researchers discovered a new type of ransomware targeting specific individuals and distribution groups in the education and healthcare sectors. From our analysis, we noticed that managerial positions were the most frequent targets in the "RETAIL" business cluster. This concurs with the 2015 "The Human Factor" report in which security researchers found that middle management targets were on the increase due to the likelihood of these individuals opening malicious emails (Proofpoint, 2015)

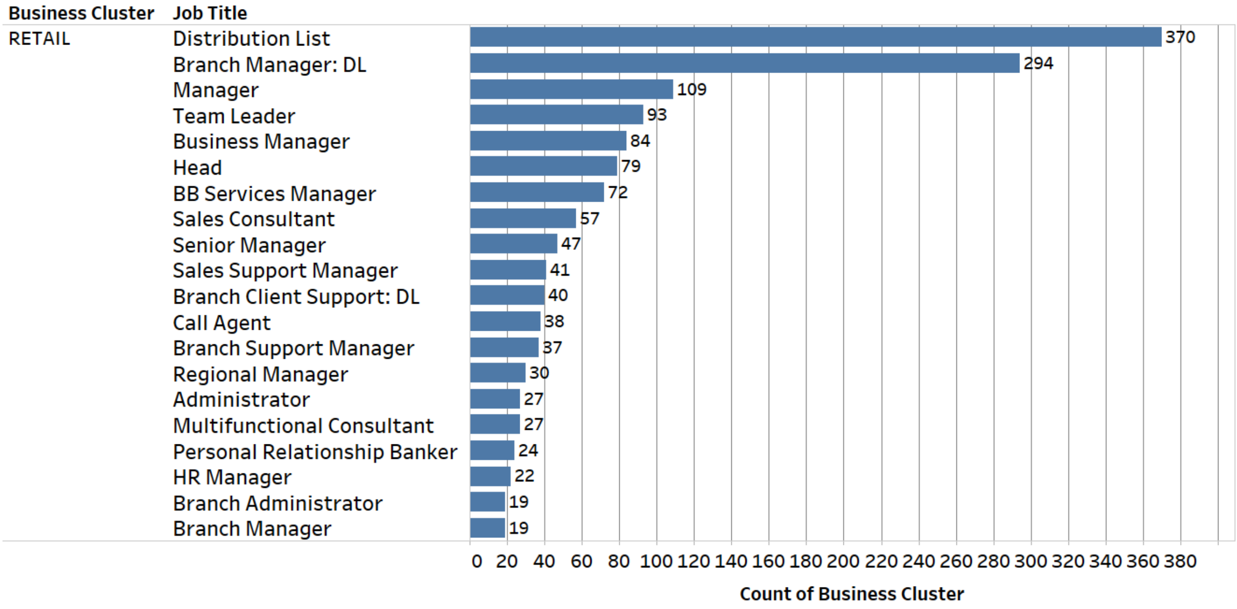


Figure 4.12: Top 20 job titles targeted in retail

4.3.2 Top 10 targeted departments in the bank

Owing to the widespread distribution of different departments in the bank, this particular attribute did not provide much value, as shown in Figure 4.13. One noticeable finding that correlates with the business cluster discussion in Section 4.3.1 was the number of invalid email addresses targeted.

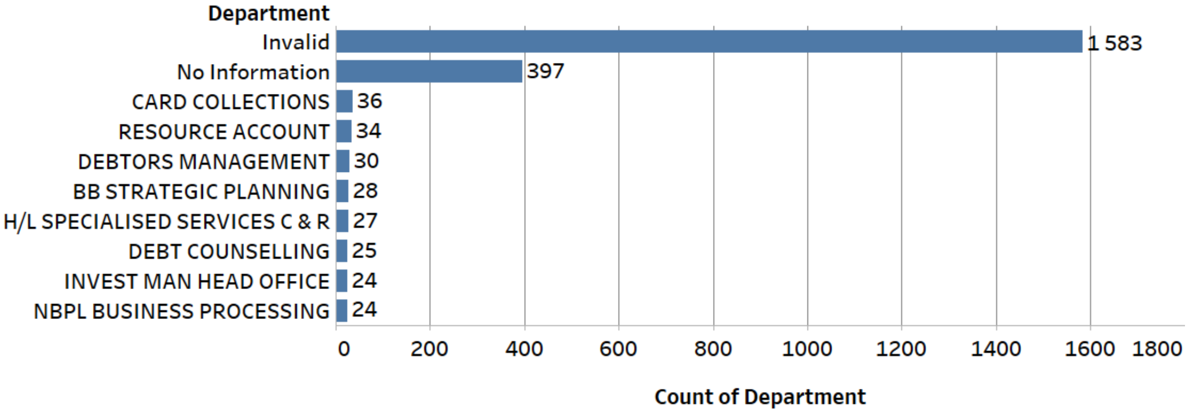
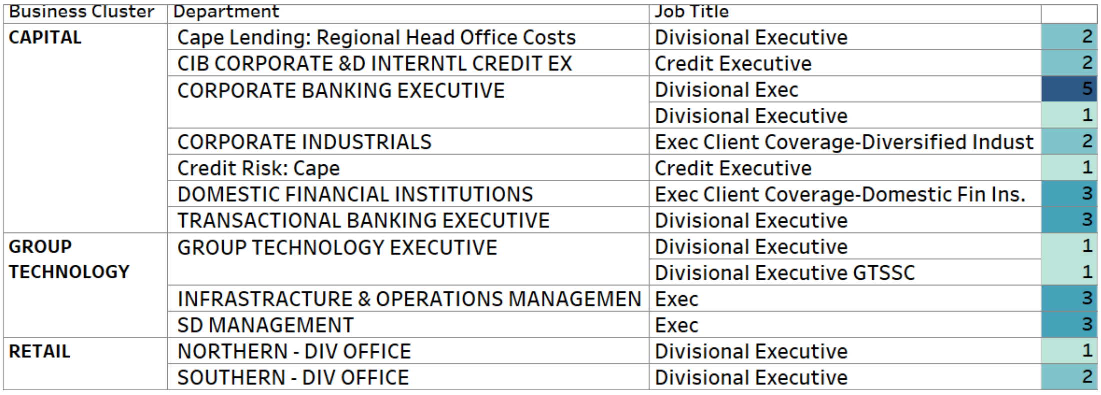


Figure 4.13: Top 10 targeted departments in the bank

However, the department attribute can still prove valuable when used in conjunction with the business cluster and job title attributes. For example, considering the executive leadership job title identified by the presence of the string “Exec” in the job title field, we noticed that the “Divisional Exec” job title in the “Corporate Banking Executive” department within the “Capital” business cluster was the most targeted during the 6 month observation period. Therefore, the department attribute could still provide valuable information when used on a granular level as shown in Figure 4.14.



Business Cluster	Department	Job Title	
CAPITAL	Cape Lending: Regional Head Office Costs	Divisional Executive	2
	CIB CORPORATE &D INTERNTL CREDIT EX	Credit Executive	2
	CORPORATE BANKING EXECUTIVE	Divisional Exec	5
		Divisional Executive	1
	CORPORATE INDUSTRIALS	Exec Client Coverage-Diversified Indust	2
	Credit Risk: Cape	Credit Executive	1
	DOMESTIC FINANCIAL INSTITUTIONS	Exec Client Coverage-Domestic Fin Ins.	3
	TRANSACTIONAL BANKING EXECUTIVE	Divisional Executive	3
GROUP TECHNOLOGY	GROUP TECHNOLOGY EXECUTIVE	Divisional Executive	1
		Divisional Executive GTSSC	1
	INFRASTRUCTURE & OPERATIONS MANAGEMEN	Exec	3
	SD MANAGEMENT	Exec	3
RETAIL	NORTHERN - DIV OFFICE	Divisional Executive	1
	SOUTHERN - DIV OFFICE	Divisional Executive	2

Figure 4.14: Executive leadership positions targeted

4.3.3 Top 20 targeted job titles

As shown in Figure 4.15, invalid email addresses were most targeted followed by internal distribution groups like the “Distribution List” and “Branch Manager:DL”.

Different types of managerial roles and the job title “Head” were the next most targeted. The “Head” job title represents the senior leadership layer immediately below the executive leadership.

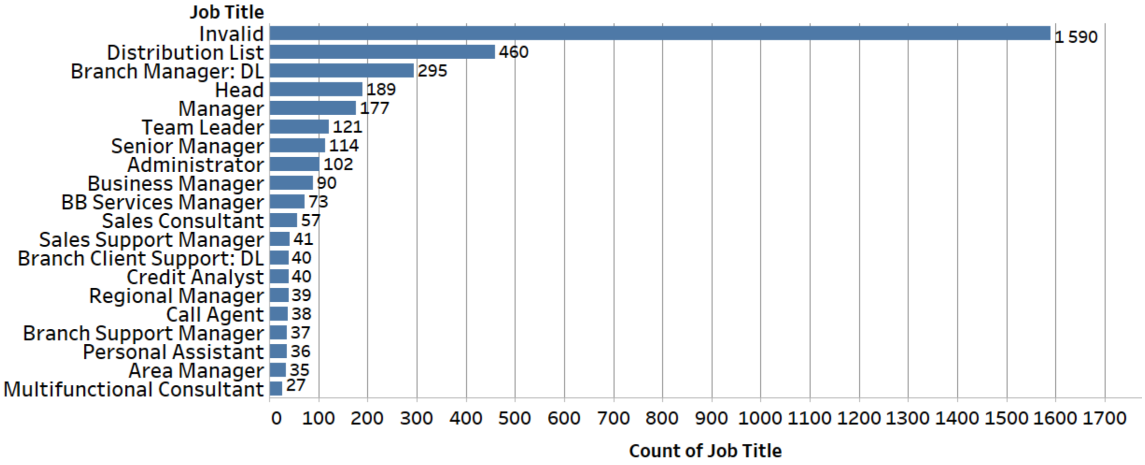


Figure 4.15: Top 20 targeted job titles

4.3.4 Summary of information from data set 3

The results of data set 3 provided a business context awareness layer. We identified the most targeted business cluster, department and job title. In addition, we obtained more granular information about executive leadership that were targeted in the bank.

4.4 Summary

In this chapter we presented the results of the individual analyses of the three different data sets. From data set 1, we identified “from where” the attacks originated, “when” the attacks occurred and “what” type of themes were used. Using data set 2, we identified “what” type of malware was used and the characteristics thereof. Finally data set 3, provided information on “who” was targeted based on the business context obtained from the AD. Thus, the results from of each data set provided a particular layer of intelligence. In Chapter 5, we discuss the added value obtained from the fused data set.

Chapter 5

Results of Fused Data Set Analysis

“The purpose of visualization is insight, not pictures.” –Ben Shneiderman

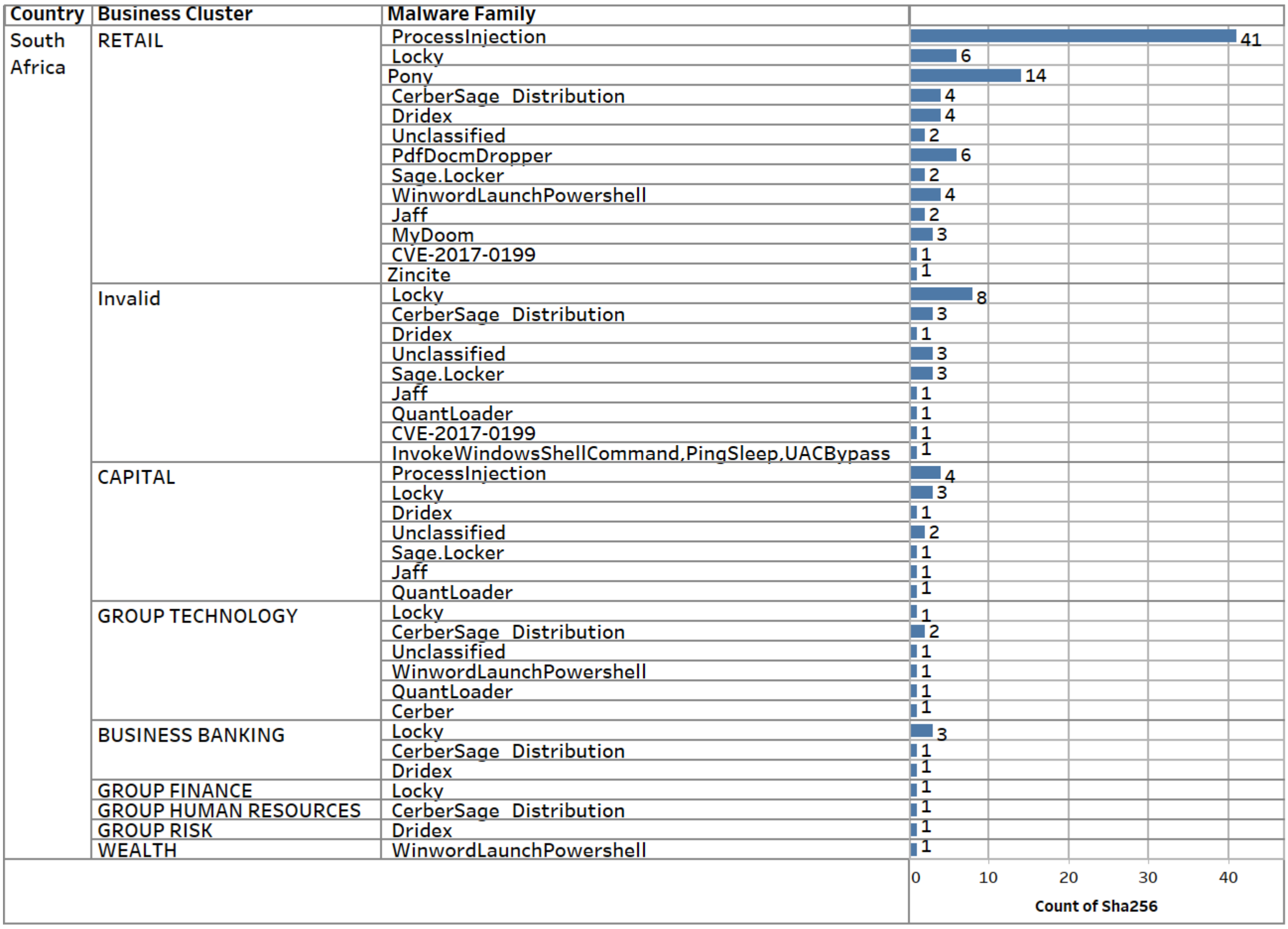
This chapter deals with the analysis of the fused data set. The researcher’s domain knowledge played a vital role in the interpretation of the fused data set. For example, the majority of malicious emails discussed in Section 4.1.1 showed constant attacks from malicious spam networks like the *Necurs* botnet. However, attack indicators such as the use of exploit code discussed in Section 4.2.2 could indicate slightly more targeted attacks based on the industry references where the use of such exploit code was reported in targeted attacks. In the analysis of the fused data set, threat scenarios are used to contextualize the value of the fused data.

For the purpose of our research, a threat scenario is defined as threat intelligence obtained through the analysis of the fused data set and therefore not as a result of the individual data set analysis as discussed in Chapter 4. Analysis of the fused data set provided a holistic view of the full email attack chain, for example, the targets (who), the time of attack (when) and the theme and malware used in the attack (what).

In the following sections we discuss the results of the analysis of the fused data set in the context of a threat scenario.

5.1 Threat scenario 1 – Business clusters targeted with malware families originating within South Africa

In Section 4.1.2 we discussed the importance of geographic situational awareness where cyber criminals tend to distribute malicious emails from within the victim's own country to evade security controls. However, malware intelligence and business context were not considered in the individual data set analysis. As the first scenario, we selected South Africa as the source country from where the malicious emails originated and expanded the targeted business clusters and the type of malware used as shown in Figure 5.1. We found that the majority of malware were distributed from malicious spam botnets like the *Necurs* botnet as discussed in Section 4.2, thereby indicating that malicious distribution networks made use of within country malicious email distribution to evade potential geographic restrictions. However, the unclassified, behavioural attributes and exploit code of the malware should be investigated further to determine if these formed part of malicious spam botnet attacks or targeted attacks.



Country	Business Cluster	Malware Family	Count of Sha256
South Africa	RETAIL	ProcessInjection	41
		Locky	6
		Pony	14
		CerberSage_Distribution	4
		Dridex	4
		Unclassified	2
		PdfDocmDropper	6
		Sage.Locker	2
		WinwordLaunchPowershell	4
		Jaff	2
		MyDoom	3
		CVE-2017-0199	1
		Zincite	1
	Invalid	Locky	8
		CerberSage_Distribution	3
		Dridex	1
		Unclassified	3
		Sage.Locker	3
		Jaff	1
		QuantLoader	1
		CVE-2017-0199	1
		InvokeWindowsShellCommand,PingSleep,UACBypass	1
	CAPITAL	ProcessInjection	4
		Locky	3
		Dridex	1
		Unclassified	2
		Sage.Locker	1
		Jaff	1
		QuantLoader	1
	GROUP TECHNOLOGY	Locky	1
		CerberSage_Distribution	2
		Unclassified	1
		WinwordLaunchPowershell	1
		QuantLoader	1
		Cerber	1
	BUSINESS BANKING	Locky	3
		CerberSage_Distribution	1
		Dridex	1
	GROUP FINANCE	Locky	1
	GROUP HUMAN RESOURCES	CerberSage_Distribution	1
	GROUP RISK	Dridex	1
	WEALTH	WinwordLaunchPowershell	1

Figure 5.1: Business clusters targeted with malware originating from within South Africa

5.2 Threat scenario 2 – Subject line analysis of *Locky* ransomware from the top 5 countries

Referring to Section 4.2.1, the *Locky* ransomware was one of the top malware families that targeted the bank through malicious spam networks like the *Necurs* botnet. In threat scenario 2, we selected the *Locky* ransomware, the subject lines used by *Locky* and the top 5 countries from where it originated. India accounted for the majority of the *Locky* ransomware distribution. In addition, we observed a particular order of subject lines used in the distribution of *Locky* from the top 5 countries as shown in Figure 5.2.

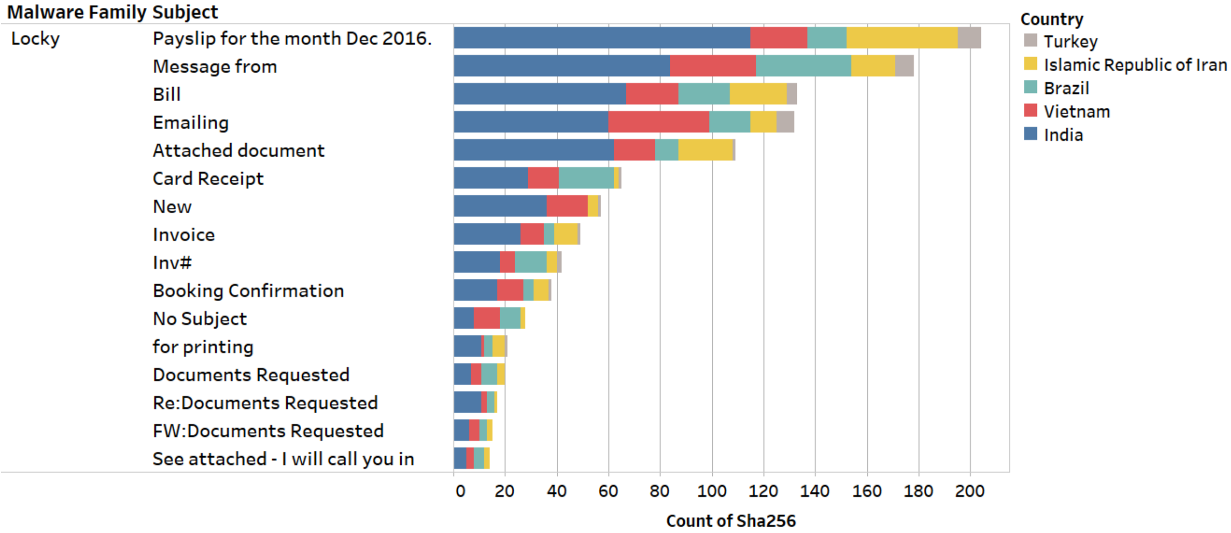
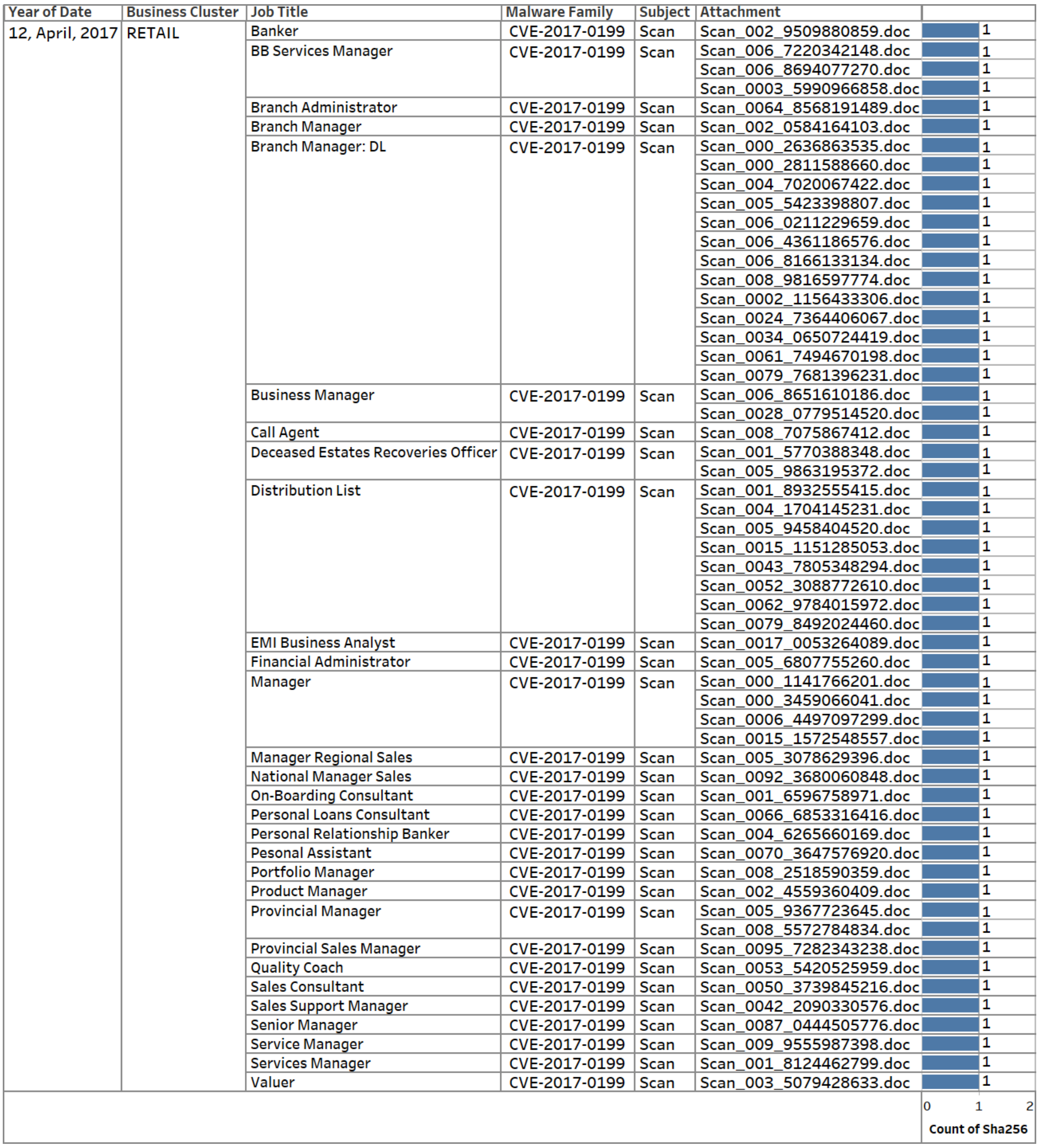


Figure 5.2: Subject line analysis of *Locky* ransomware from the top 5 countries

5.3 Threat scenario 3 – Distribution analysis of exploit code "CVE-2017-0199"

In threat scenario 3, we expanded on the exploit code "CVE-2017-0199" discussed in Section 4.2.2. On April 11 2017, Jiang *et al.* (2017) reported that exploit code targeting a Microsoft rich text format (RTF) document vulnerability known as "CVE-2017-0199" was detected in targeted attacks against organizations. Szappanos (2017) also reported that sophisticated cyber criminals made use of the "CVE-2017-0199" exploit code in attacks prior to April 2017.

The "CVE-2017-0199" exploit code was selected and filtered on the date, "RETAIL" business cluster, job title, subject and attachment. According to our results, the exploit code for "CVE-2017-0199" was used in an attack on the bank on April 12 2017, only one day after Jiang *et al.* (2017) reported the use of such exploit code as shown in Figure 5.3. However, based on the character randomization of the attachment field and the use of the same subject line, it would seem plausible that the attack originated from a malicious spam botnet (Magnúsardóttir, 2017). In addition, Szappanos (2017) reported that malicious spam networks made use of the "CVE-2017-0199" exploit code to distribute financial malware like *Dridex*.

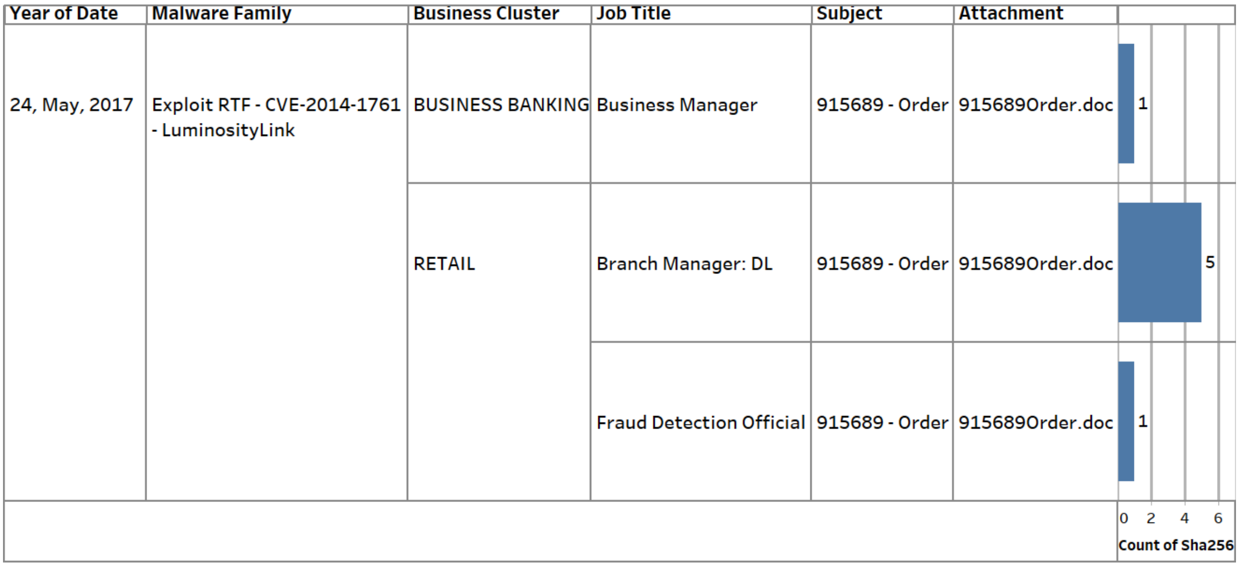


Year of Date	Business Cluster	Job Title	Malware Family	Subject	Attachment	
12, April, 2017	RETAIL	Banker	CVE-2017-0199	Scan	Scan_002_9509880859.doc	1
		BB Services Manager	CVE-2017-0199	Scan	Scan_006_7220342148.doc	1
					Scan_006_8694077270.doc	1
					Scan_0003_5990966858.doc	1
		Branch Administrator	CVE-2017-0199	Scan	Scan_0064_8568191489.doc	1
		Branch Manager	CVE-2017-0199	Scan	Scan_002_0584164103.doc	1
		Branch Manager: DL	CVE-2017-0199	Scan	Scan_000_2636863535.doc	1
					Scan_000_2811588660.doc	1
					Scan_004_7020067422.doc	1
					Scan_005_5423398807.doc	1
					Scan_006_0211229659.doc	1
					Scan_006_4361186576.doc	1
					Scan_006_8166133134.doc	1
					Scan_008_9816597774.doc	1
					Scan_0002_1156433306.doc	1
					Scan_0024_7364406067.doc	1
					Scan_0034_0650724419.doc	1
					Scan_0061_7494670198.doc	1
					Scan_0079_7681396231.doc	1
		Business Manager	CVE-2017-0199	Scan	Scan_006_8651610186.doc	1
					Scan_0028_0779514520.doc	1
		Call Agent	CVE-2017-0199	Scan	Scan_008_7075867412.doc	1
		Deceased Estates Recoveries Officer	CVE-2017-0199	Scan	Scan_001_5770388348.doc	1
					Scan_005_9863195372.doc	1
		Distribution List	CVE-2017-0199	Scan	Scan_001_8932555415.doc	1
					Scan_004_1704145231.doc	1
					Scan_005_9458404520.doc	1
					Scan_0015_1151285053.doc	1
					Scan_0043_7805348294.doc	1
					Scan_0052_3088772610.doc	1
					Scan_0062_9784015972.doc	1
					Scan_0079_8492024460.doc	1
		EMI Business Analyst	CVE-2017-0199	Scan	Scan_0017_0053264089.doc	1
		Financial Administrator	CVE-2017-0199	Scan	Scan_005_6807755260.doc	1
		Manager	CVE-2017-0199	Scan	Scan_000_1141766201.doc	1
					Scan_000_3459066041.doc	1
					Scan_0006_4497097299.doc	1
					Scan_0015_1572548557.doc	1
		Manager Regional Sales	CVE-2017-0199	Scan	Scan_005_3078629396.doc	1
		National Manager Sales	CVE-2017-0199	Scan	Scan_0092_3680060848.doc	1
		On-Boarding Consultant	CVE-2017-0199	Scan	Scan_001_6596758971.doc	1
		Personal Loans Consultant	CVE-2017-0199	Scan	Scan_0066_6853316416.doc	1
		Personal Relationship Banker	CVE-2017-0199	Scan	Scan_004_6265660169.doc	1
		Pesonal Assistant	CVE-2017-0199	Scan	Scan_0070_3647576920.doc	1
		Portfolio Manager	CVE-2017-0199	Scan	Scan_008_2518590359.doc	1
		Product Manager	CVE-2017-0199	Scan	Scan_002_4559360409.doc	1
		Provincial Manager	CVE-2017-0199	Scan	Scan_005_9367723645.doc	1
					Scan_008_5572784834.doc	1
		Provincial Sales Manager	CVE-2017-0199	Scan	Scan_0095_7282343238.doc	1
		Quality Coach	CVE-2017-0199	Scan	Scan_0053_5420525959.doc	1
		Sales Consultant	CVE-2017-0199	Scan	Scan_0050_3739845216.doc	1
		Sales Support Manager	CVE-2017-0199	Scan	Scan_0042_2090330576.doc	1
		Senior Manager	CVE-2017-0199	Scan	Scan_0087_0444505776.doc	1
		Service Manager	CVE-2017-0199	Scan	Scan_009_9555987398.doc	1
		Services Manager	CVE-2017-0199	Scan	Scan_001_8124462799.doc	1
		Valuer	CVE-2017-0199	Scan	Scan_003_5079428633.doc	1

Figure 5.3: Distribution analysis of exploit code “CVE-2017-0199”

5.4 Threat scenario 4 – Distribution analysis of exploit code “CVE-2014-1761” – *LuminosityLink*

In threat scenario 4, we expanded on the exploit code “CVE-2014-1761” discussed in Section 4.2.2. In 2015, “CVE-2014-1761” was reported (Grunzweig, 2016) as the most widely exploited vulnerability by cyber criminals including Nigerian cyber criminals that targeted banks. The *LuminosityLink* malware associated with this particular exploit code allowed cyber criminals remote access to a system and key logging functionality, therefore making it a serious threat to an organization (Grunzweig, 2016). The “Exploit RTF - CVE-2014-1761 - *LuminosityLink*” exploit code was selected and filtered on the date, business cluster, job title, subject and attachment. Our results show that the “CVE-2014-1761” exploit code and associated malware were sent to particular targets in the bank as shown in Figure 5.4.

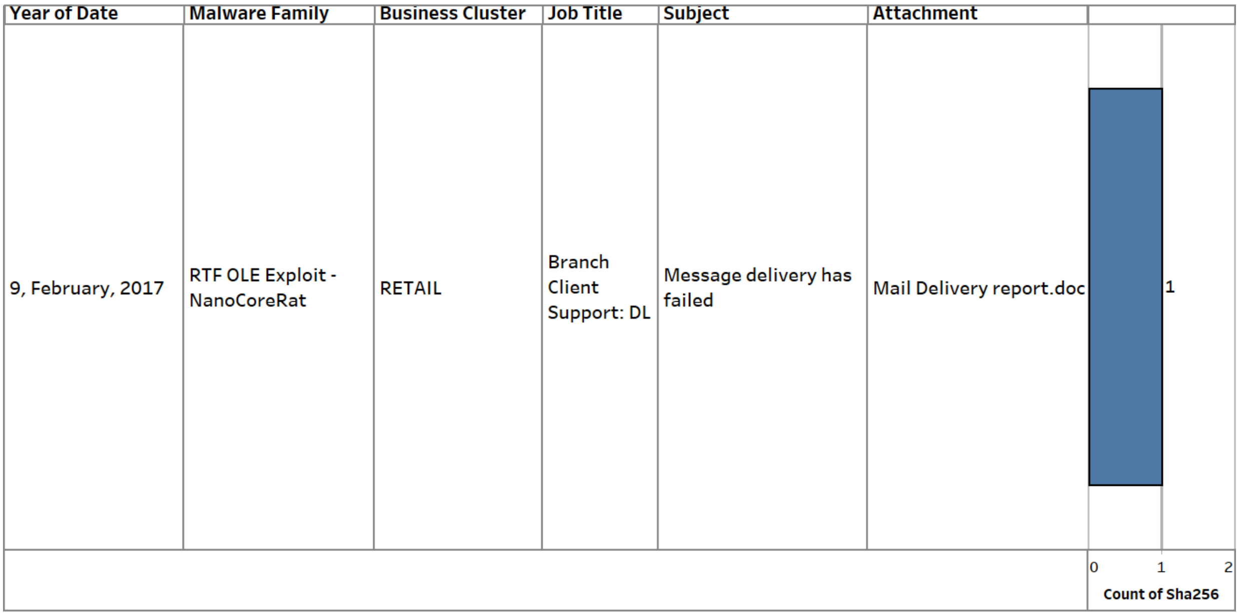


Year of Date	Malware Family	Business Cluster	Job Title	Subject	Attachment	
24, May, 2017	Exploit RTF - CVE-2014-1761 - LuminosityLink	BUSINESS BANKING	Business Manager	915689 - Order	9156890rder.doc	1
		RETAIL	Branch Manager: DL	915689 - Order	9156890rder.doc	5
			Fraud Detection Official	915689 - Order	9156890rder.doc	1

Figure 5.4: Distribution analysis of exploit code “CVE-2014-1761 – *LuminosityLink*”

5.5 Threat scenario 5 – Distribution analysis of exploit code “RTF OLE Exploit – *NanoCoreRat*”

In this scenario, we expanded on the exploit code “RTF OLE Exploit” discussed in Section 4.2.2. Although, we could not attribute this particular exploit code to a “CVE” industry identifier, the *NanoCoreRat* malware associated with this attack has been attributed to targeted attacks dating back to 2015 as reported by Llascu (2015) and the Talos Group (2015). The *NanoCoreRat* malware allowed cyber criminals remote access to a network, therefore making it a serious threat to an organization. The “RTF OLE Exploit - *NanoCoreRat*” exploit code was selected and filtered on the date, business cluster, job title, subject and attachment. In the fused data set, our data showed that a single distribution group in the “RETAIL” business cluster was targeted as shown in Figure 5.5.

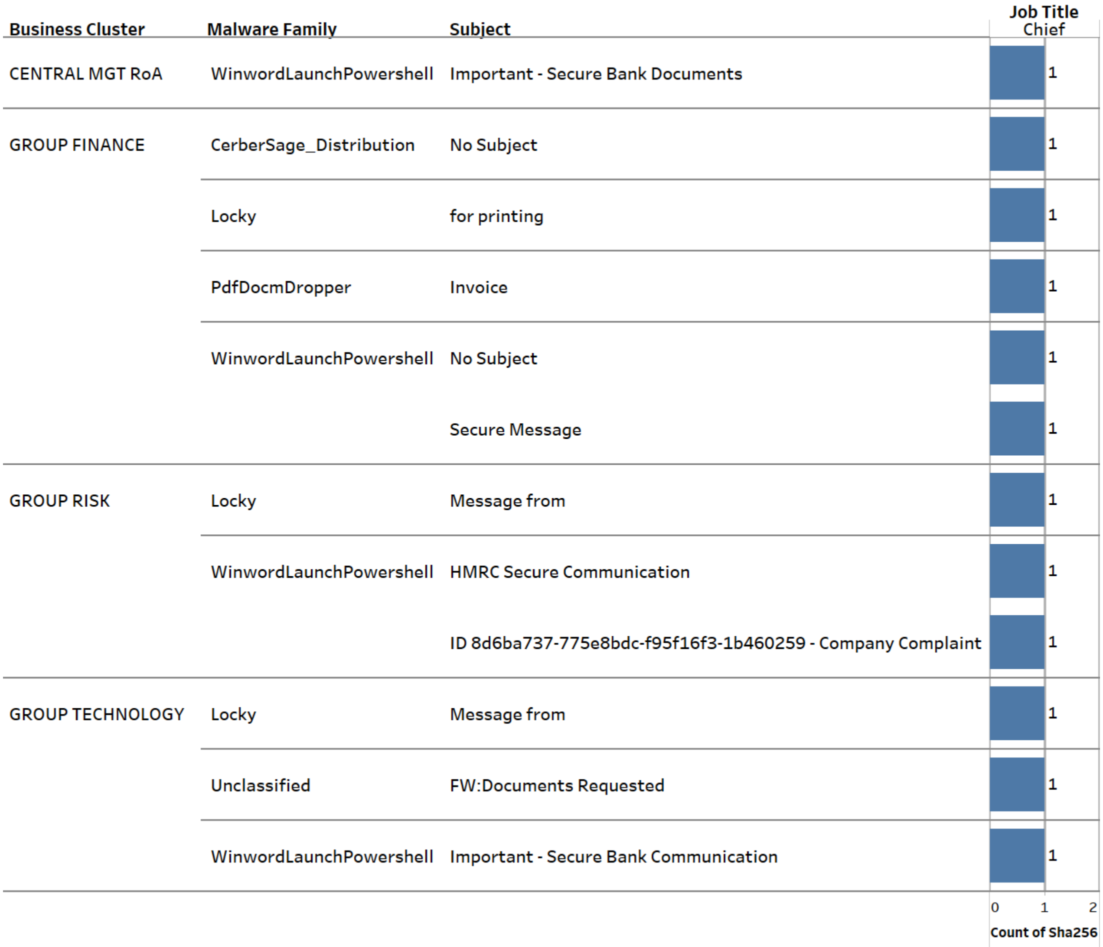


Year of Date	Malware Family	Business Cluster	Job Title	Subject	Attachment
9, February, 2017	RTF OLE Exploit - NanoCoreRat	RETAIL	Branch Client Support: DL	Message delivery has failed	Mail Delivery report.doc

Figure 5.5: Distribution analysis of exploit code “RTF OLE Exploit – *NanoCoreRat*”

5.6 Threat scenario 6 – Malware distribution analysis of C-level targets in the bank

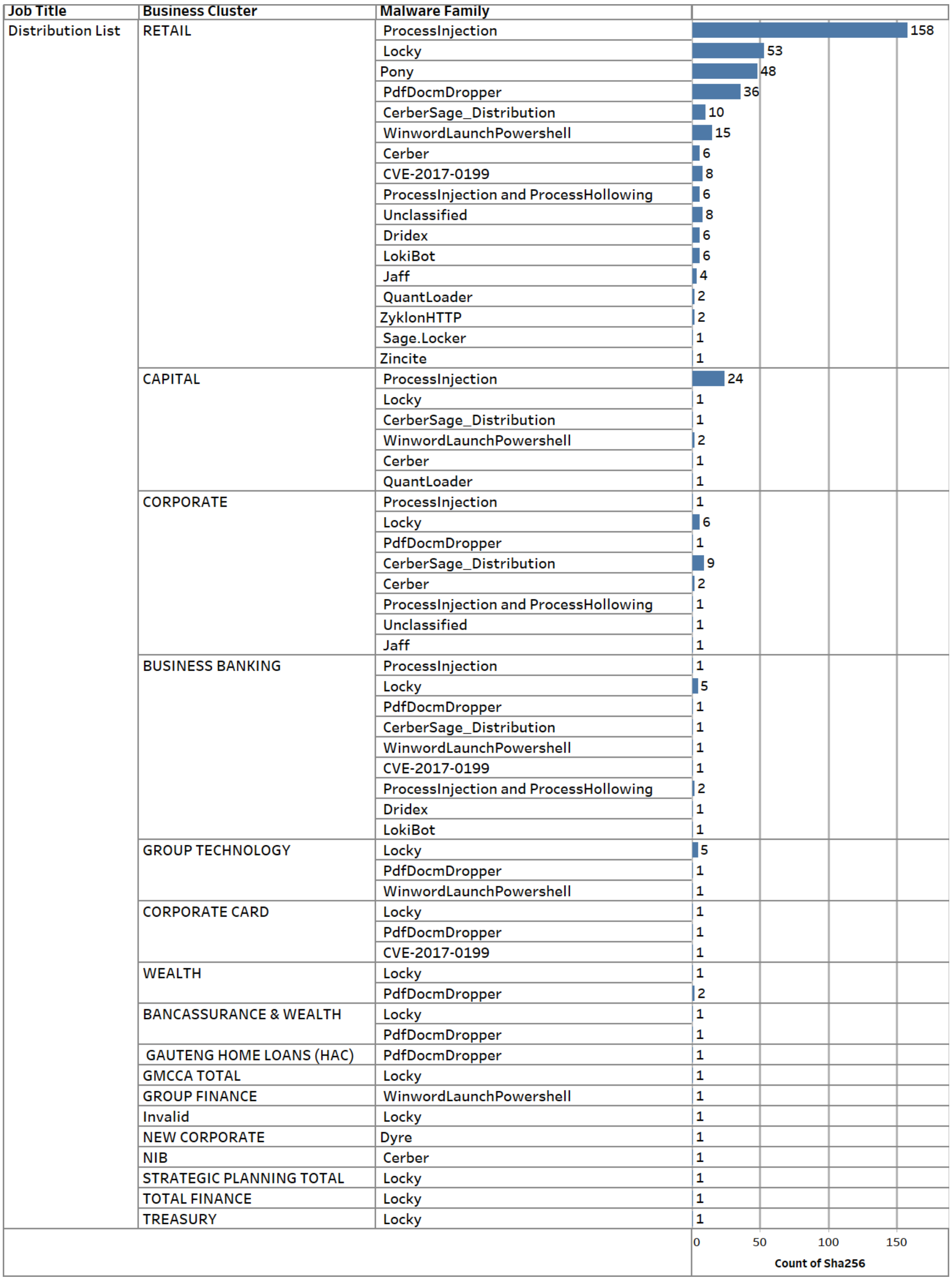
In this scenario, we selected the “Chief” job title, which represents the highest ranking job title of an individual in the bank, and filtered on the business cluster, malware family and subject. In our results, we found malicious spam botnet activity targeting “Chief” job titles across the different business clusters, followed by suspicious PowerShell activity across three of the business clusters and a unique unclassified malware sample targeting a “Chief” job title in “GROUP TECHNOLOGY” as shown in Figure 5.6.



Business Cluster	Malware Family	Subject	Job Title Chief
CENTRAL MGT RoA	WinwordLaunchPowershell	Important - Secure Bank Documents	1
GROUP FINANCE	CerberSage_Distribution	No Subject	1
	Locky	for printing	1
	PdfDocmDropper	Invoice	1
	WinwordLaunchPowershell	No Subject	1
		Secure Message	1
GROUP RISK	Locky	Message from	1
	WinwordLaunchPowershell	HMRC Secure Communication	1
		ID 8d6ba737-775e8bdc-f95f16f3-1b460259 - Company Complaint	1
GROUP TECHNOLOGY	Locky	Message from	1
	Unclassified	FW:Documents Requested	1
	WinwordLaunchPowershell	Important - Secure Bank Communication	1

Figure 5.6: Malware distribution analysis of C-level targets in the bank

5.7 Threat scenario 7 – Malware distribution analysis of the most targeted job title



Job Title	Business Cluster	Malware Family	
Distribution List	RETAIL	ProcessInjection	158
		Locky	53
		Pony	48
		PdfDocmDropper	36
		CerberSage_Distribution	10
		WinwordLaunchPowershell	15
		Cerber	6
		CVE-2017-0199	8
		ProcessInjection and ProcessHollowing	6
		Unclassified	8
		Dridex	6
		LokiBot	6
		Jaff	4
		QuantLoader	2
		ZyklonHTTP	2
		Sage.Locker	1
		Zincite	1
	CAPITAL	ProcessInjection	24
		Locky	1
		CerberSage_Distribution	1
		WinwordLaunchPowershell	2
		Cerber	1
		QuantLoader	1
	CORPORATE	ProcessInjection	1
		Locky	6
		PdfDocmDropper	1
		CerberSage_Distribution	9
		Cerber	2
		ProcessInjection and ProcessHollowing	1
		Unclassified	1
		Jaff	1
	BUSINESS BANKING	ProcessInjection	1
		Locky	5
		PdfDocmDropper	1
		CerberSage_Distribution	1
		WinwordLaunchPowershell	1
		CVE-2017-0199	1
		ProcessInjection and ProcessHollowing	2
		Dridex	1
		LokiBot	1
	GROUP TECHNOLOGY	Locky	5
		PdfDocmDropper	1
		WinwordLaunchPowershell	1
	CORPORATE CARD	Locky	1
		PdfDocmDropper	1
		CVE-2017-0199	1
	WEALTH	Locky	1
		PdfDocmDropper	2
	BANCASSURANCE & WEALTH	Locky	1
		PdfDocmDropper	1
	GAUTENG HOME LOANS (HAC)	PdfDocmDropper	1
	GMCCA TOTAL	Locky	1
	GROUP FINANCE	WinwordLaunchPowershell	1
	Invalid	Locky	1
	NEW CORPORATE	Dyre	1
	NIB	Cerber	1
	STRATEGIC PLANNING TOTAL	Locky	1
	TOTAL FINANCE	Locky	1
	TREASURY	Locky	1

Figure 5.7: Malware distribution analysis of the most targeted job title

In Section 4.3.1 we discussed the most targeted job title “Distribution List”, which received a total of 460 malicious emails. In this threat scenario, the job title “Distribution List” was selected and filtered by business cluster and malware family. Our results showed that only one out of the 460 “Distribution list” targets were invalid and that a variety of different malware families were used in targeting distribution groups as shown in Figure 5.7.

5.8 Summary

In this chapter we presented the results of the fused data set analysis by means of seven threat scenarios. The fused data set provided the context to answer the “who”, “what”, “where” and “when” instead of performing lengthy fragmented analyses across three different data sets. In threat scenarios 4 and 5, we observed indicators of more targeted attacks than the usual malicious spam observed. Therefore, the analysis of the fused data set provided a holistic view of malicious email targeting the bank as well as security awareness of the types of malware used in the attacks and who the targets were. Many different threat scenarios could be developed to answer different business related questions about malicious email attacks on the bank. However, for the purpose of this research, the seven threat scenarios were considered to be sufficient to illustrate the value of the fused data set. In the following chapter we discuss the usefulness of the fused data set results.

Chapter 6

Relevance of Information from the Fused Data Set

"There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns — the ones we don't know we don't know"
–Donald Rumsfeld

Threat intelligence often includes many different types of data sources. For example, perimeter sensor data, network traffic and other security information, and event management data. The data are usually stored and analysed separately. Often, the fragmented analysis of such data does not provide the required value to the business. In Section 6.2 we discuss the limitations of each data set, followed by a discussion of the fused data set in Section 6.3. Finally, we discuss the relevance of the fused data set results in Section 6.4.

6.1 Limitations of each data set

To understand the relevance of the fused data set results, the limitations of each data set must first be determined.

6.1.1 Limitations of data set 1

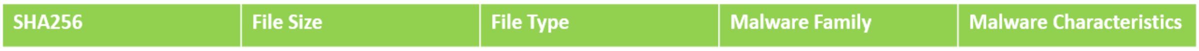
In data set 1, a malicious email attachment was detected on the bank's AMEPS. The AMEPS log data provided the time and date it was sent, the sender email and IP address, the recipient email address, the subject and attachment, the SHA256 cryptographic identifier associated with the malicious email and the country it was sent from as shown in Table 6.1. However, using data set 1, we could not ascertain the type or characteristics of the malware, thereby limiting our understanding of the functionality and purpose of the malware. Furthermore, the recipient email address did not provide the required business context about who the target was in the organization.

Table 6.1: Information available from data set 1

Time	Date	Protocol	Sender IP	Sender	Recipient	Subject	Attachment	SHA256	Geo Location	Status

6.1.2 Limitations of data set 2

Using data set 2, information of the malware family was obtained based on the SHA256 cryptographic identifier identified in data set 1. In the absence of the malware family, we obtained characteristics of the malware and in the absence of the malware family and characteristics, we labeled the malware as "Unclassified". In addition, we obtained the file size and file type of the malware as shown in Figure 6.1. However, data set 2 could not provide any business context about the target. Furthermore, without the SHA256 cryptographic identifier values obtained in data set 1, it would not have been possible to identify the malware family and/or characteristics of the malware.

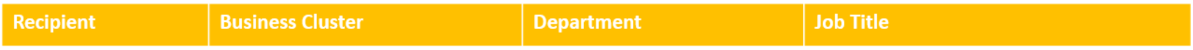


SHA256	File Size	File Type	Malware Family	Malware Characteristics

Figure 6.1: Information available from data set 2

6.1.3 Limitations of data set 3

In data set 3, the business context based on the recipient email address that was targeted was extracted from the AD. For example, the business cluster, department and job title of the target was extracted as shown in Figure 6.2. However, the business context data did not provide the malware intelligence, where the mail originated from or the theme of the malicious email. Furthermore, without the recipient email address obtained from data set 1, it would not have been possible to extract the business context from AD.

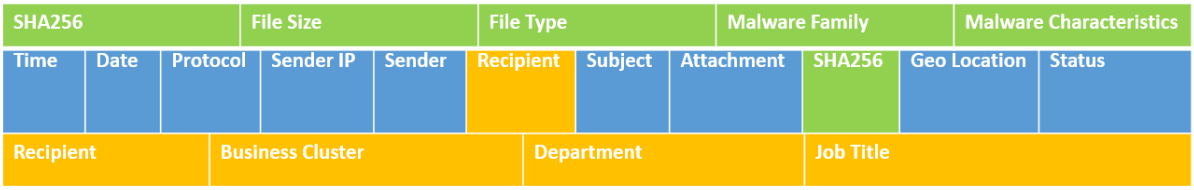


Recipient	Business Cluster	Department	Job Title

Figure 6.2: Information available from data set 3

6.2 Fused data set

In the fused data set all three data sets were joined based on a common field. Data sets 1 and 2 were joined based on the SHA256 cryptographic identifier to expand data set 1 with the malware intelligence data. Data sets 1 and 3 were joined based on the recipient field to expand data sets 1 and 2 with the business context data from the AD as shown in Figure 6.3.



SHA256	File Size	File Type	Malware Family	Malware Characteristics

Time	Date	Protocol	Sender IP	Sender	Recipient	Subject	Attachment	SHA256	Geo Location	Status

Recipient	Business Cluster	Department	Job Title

Figure 6.3: Information available from the fused data set

6.3 Relevance of the fused data set information

The information obtained from the fused data set provided a workable level of situational awareness and threat intelligence data that could be used in proactive cyber defence. In the following sections we discuss the relevance of the fused data set information within a threat intelligence context as discussed in Section 2.6.

6.3.1 Technical threat intelligence

The SHA 256 cryptographic identifiers associated with the malicious emails sent to the bank were extracted from data set 1, thereby providing the malware hashes considered as static data points in technical threat intelligence as discussed in Section 2.6.2.

6.3.2 Tactical threat intelligence

The malware hashes obtained from data set 1 were used to obtain malware intelligence, which was correlated with industry research as discussed in Section 4.2. This provided TTPs that could be used by incident response and monitoring teams. For example, malware attributes such as the use of PowerShell, process injection and/or process hollowing indicated the tools and techniques used by the malware. Therefore, the fusion of data set 2 provided tactical threat intelligence as discussed in Section 2.6.2.

6.3.3 Operational threat intelligence

In the most uncomplicated way, situational awareness means being aware of what is happening around you (Varga *et al.*, 2016). Data set 1 provided the basic information about the malicious email attachments used in email attacks against the bank. Therefore, we were aware that the bank was being targeted by malware. However, we had no context of the type of malware used and/or the business context of the victims targeted in the bank. The fused data set discussed in Chapter 5, provided an extra layer of abstraction about the type of malware used and who the targets were. For example, in threat scenario 4 discussed in Section 5.4, a branch manager, fraud detection official and branch manager distribution list were targeted with exploit code using a known malware family reportedly used in targeted attacks on organizations previously. In Section 4.2.2 we discussed the use of the "CVE-2017-0199" exploit code in targeted and malicious spam attacks. The awareness of such exploit code being used in the attacks provided actionable intelligence that could be used to determine the bank's level of exposure against such attacks. For example, the use of the "CVE-2017-0199" exploit code within 24 hours after Microsoft released the software update would require a concerted effort to apply the required software updates to mitigate the threat. In the absence of applying the required software updates in time, the bank could review the current security controls and implement preventative measures whilst the software update processes were underway.

The majority of malicious emails targeted internal distribution lists used across the different business clusters in the bank of which only one was found to be invalid. The internal distribution lists targeted were linked to the business clusters targeted and the different types of malware used. In Section 4.3.1 it was stated that attackers often target internal distribution lists due to the relaxed security controls and the wider reach it provided into an organization. This suggested another layer of actionable intelligence. For example, the bank could review the security controls in place on the internal distribution lists targeted. In addition, security awareness training could be focused on the employees receiving email from such distribution lists. The results of the fused data set provided deeper insight into the email attacks experienced by the bank. Therefore, situational awareness was obtained from the fused data set, which in turn provided operational threat intelligence that could be used to improve security controls as discussed in Section 2.6.2.

6.3.4 Strategic threat intelligence

The fused data set provided insight about the volume of ransomware attacks experienced in the "RETAIL" business cluster. This information could be used to determine the potential financial impact on the bank and in particular the "RETAIL" business cluster should such an attack be successful. This contributed to strategic threat intelligence in terms of determining the potential financial impact of a threat as discussed in Section 2.6.2.

6.4 Summary

In this chapter we discussed the relevance of the information obtained from the fused data set. This information provided actionable intelligence that could be linked to each of the four sub areas of threat intelligence as discussed in Section 2.6.2. Therefore, the analysis of fusing advanced malware email protection logs, malware intelligence and AD attributes could be used as an instrument for threat intelligence.

Chapter 7

Conclusion and Future Work

”To be suspicious is not a fault. To be suspicious all the time without coming to a conclusion is the defect.” –Lu Xun

The last chapter outlines the primary findings and highlights the contribution of the research. The chapter closes by recommending future research work that can be built on this research.

7.1 Summary of Research

With email being one of the largest electronic communication mediums used today, it could be argued that malware is one of the most significant cyber threats an organization faces daily as cyber criminals prey on unsuspecting victims opening malicious email attachments.

As organizations make use of the latest advanced malware protection systems to combat such threats, understanding the type of malware used and the victims targeted is becoming more important holistically. Malware laced emails could form part of automated opportunistic attempts at attacks or part of a sophisticated targeted attack chain. Saying that, in a recent report by RSA (2017), it was reported that the *Carbanak* threat actor discussed in the literature review, made use of infrastructure used by opportunistic threat actors, thereby harvesting potential information of compromised targets to focus their operations on.

In 2017, developing threat intelligence from internal data log sources surfaced as a trend between cyber security teams. However, deriving threat intelligence from log data is not that straight forward and needs to go through a cycle to determine the relevance.

This research confirmed that making use of extensive data fusion techniques on log data could be used as a source of the four areas of threat intelligence. In addition, this research provides a working methodology that could be applied to different malware data sets.

7.2 Contributions of Research

All the research objectives set out in Chapter 1 have been realised in this thesis and are revisited below, along with a discussion on the degree to which they have been achieved.

1. *To what extent can the fusion of AMP logs, malware intelligence and AD user attributes be used as an instrument for threat intelligence?*

   This first objective has been met. As stated in Section 2.6 threat intelligence can aid decision makers on how to respond to threats. As an example, the use of exploit code discussed in the threat scenarios in Chapter 5 provided actionable intelligence that the bank could use to determine its exposure to such threats and prioritize the remediation thereof as discussed in the relevance of the fused data set information in Section 6.3.3.

2. *To what extent can the fusion of AMP logs, malware intelligence and AD user attributes be used in a cyber security awareness program?*

   In Chapter 4, we identified the most targeted business cluster and job title, thereby identifying the most targeted business areas and internal distribution groups. Furthermore, we examined the top 20 subject lines and the uniqueness thereof. Thus, all this information could be used in a cyber security awareness program targeting those business areas and distribution groups, thereby making them aware of the malware threats and the type of subject lines used in these attacks.

3. *What type of threat intelligence can be derived by the fusion of AMP logs with AD user attributes in an organization?*

   In Section 2.6.3, we discussed the different types of threat intelligence. In Chapter 6 we determined that all four types of threat intelligence could be obtained from the fusion of AMP logs, malware intelligence and AD user attributes, thereby meeting this objective.

4. *Can the information derived from the above research questions be factored into a cyber security dashboard for executive management?*

   In Section 2.6.3, we discussed strategic threat intelligence that provides a view to executive management typically about the financial impact of threats to an organization. In Chapter 4, we identified that ransomware was one of the top malware threats targeting the bank's "RETAIL" business cluster and in Chapter 6, we determined that this information could be used to determine the financial impact of such an attack should it be successful.

7.3 Future Research and Recommendations

This section considers possible further research in the application of data fusion in advanced malware log data, malware intelligence and AD attributes.

1. Further research is warranted taking into consideration data points of an organizational human resource database. This research did not factor in the time and date of employees leaving the company which could be used as another data point to determine the valid period when employees were targeted with malware.

2. Another interesting data point to determine would be if email addresses were scraped or exposed through a third party data breach. For example, an organization could make use of a free service like "haveibeenpwned[1]" to verify if an email address was exposed through a data breach and use tools like "EmailHarvestor[2]" to scrape organizational email addresses to compare with the targeted email addresses.

[1] https://haveibeenpwned.com/

[2] https://github.com/maldevel/EmailHarvester

3. The researcher made use of a paid API key to use the *AutoLenz* script to obtain malware intelligence. Further research could explore open-source malware intelligence data fusion based on the SHA256 cryptographic identifiers provided in the web link that could be found in Appendix A.

4. This research was conducted utilizing static data log exports to fuse and obtain threat intelligence. Future research could explore an automated system to perform dynamic analysis of such data.

5. This research could also be expanded further by analysing the network and DNS traffic of the malware communication channels. Furthermore, network and DNS communication channels could be clustered together and mapped to the malware families, thereby providing a richer threat intelligence layer.

6. Research into threat intelligence metrics could also be considered based on this research.

7. One of the research objectives was to use the results of the fused data set in a cyber security awareness program. Further research could extend this concept by simulating similar phishing themes and/or techniques on a test group of users that have been trained with the cyber security awareness data, thereby evaluating the effectiveness of using such data in a cyber security awareness program.

It could be a matter of days, months or years before an organization is breached. Processing the internal log data of critical internal defences and fusing such data with business context can often provide early indicators to strengthen the defences against the adversary.

References

Abraham, Subil, & Nair, Suku. 2015. A Predictive Framework for Cyber Security Analytics Using Attack Graphs. *International Journal of Computer Networks & Communications (IJCNC)*, **7**(1), 1–17.

Alfreds, Duncan. 2016a. Here's how ransomware hits SA. Online. http://www.fin24.com/Tech/Cyber-Security/heres-how-ransomware-hits-sa-20160414 (accessed on 2017-04-07).

Alfreds, Duncan. 2016b. How ransomware has cost some Fin24 users thousands. Online. http://www.fin24.com/Tech/Opinion/how-ransomware-has-cost-some-fin24-users-thousands-20160415 (accessed on 2017-04-07).

Altares, Eduardo. 2017. Sage 2.0 ransomware delivered by Pandex spambot, mimics Cerber routines. Online. https://www.symantec.com/connect/blogs/sage-20-ransomware-delivered-pandex-spambot-mimics-cerber-routines (accessed on 2017-09-11).

Anderson, Vicki. 2016. Ransomware: Latest cyber extortion tool. Online. https://www.fbi.gov/contact-us/field-offices/cleveland/news/press-releases/ransomware-latest-cyber-extortion-tool (accessed on 2017-04-07).

Arnao, Matthew, Smutz, Charles, Zollman, Adam, Richardson, Andrew, & Hutchins, Eric. 2015. *Laika BOSS : Scalable File-Centric Malware Analysis and Intrusion Detection System Design*. Technical report. Lockheed Martin.

Avast. 2014. Reveton ransomware has dangerously evolved. Online. https://blog.avast.com/2014/08/19/reveton-ransomware-has-dangerously-evolved/ (accessed on 24/10/2017).

Baird, Sean, Brumaghin, Edmund, Carter, Earl, & Schultz, Jaeso. 2017. Necurs diversifies. Online. http://blog.talosintelligence.com/2017/03/necurs-diversifies.html (accessed on 2017-04-07).

Bartholomew, Brian, & Guerrero-Saade, Juan Andres. 2016. Wave Your False Flags! Deception Tactics Muddying Attribution in Targeted Attacks. *Pages 1–11 of: Virus Bulletin Conference 2016*.

Bencsáth, Boldizsár, Pék, Gábor, Buttyán, Levente, & Félegyházi, Márk. 2012. The Cousins of Stuxnet: Duqu, Flame, and Gauss. *Future Internet*, **4**(4), 971–1003.

Bilen, Tom. 2015. Using the Cloud to Minimize Security Risks. *Pages 1–55 of: Information Systems Security Association*.

Bronk, Christopher, & Tikk-Ringas, Eneken. 2013. Hack or Attack? Shamoon and the Evolution of Cyber Conflict. *Survival, Global Politics and Strategy*, 1–30.

Burbage, Paul, & Kremez, Vitali. 2017. Necurs botnet fuels massive spam campaigns spreading Jaff Ransomware. Online. `https://www.flashpoint-intel.com/blog/necurs-botnet-jaff-ransomware/` (accessed on 2017-09-09).

Chen, Hsinchun, & Storey, Veda C. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *Mis Quarterly*, **36**(4), 1165–1188.

Cheng, Rocky. 2017. Technology Risk Management in Banking Industry. *Pages 1–19 of: ISACA Annual Conference Hong Kong*.

Chismon, David, & Ruks, Martyn. 2015. *Threat Intelligence: Collecting, Analysing, Evaluating*. Technical report. MWR InfoSecurity.

Cisco. 2015. *Cisco Advanced Malware Protection*. Technical report. Cisco.

Cove, Brett. 2010. Targeted webmail phishing attacks. Online. `https://nakedsecurity.sophos.com/2010/12/16/targeted-webmail-phishing-attacks/` (accessed on 2017-09-20).

Cox, Nicholas J, & Jones, Kelvyn. 1981. Exploratory data analysis. *Quantitative Geography, London: Routledge*, 135–143.

Crest. 2015. *Cyber Security Monitoring Guide*. Technical report. CREST.

Davison, Jason. 2017. Smoke loader adds additional obfuscation methods to mitigate analysis. Online. `https://info.phishlabs.com/blog/smoke-loader-adds-additional-obfuscation-methods-to-mitigate-analysis` (accessed on 2017-09-12).

Dhamija, Rachna, & Hearst, Marti. 2006. Why Phishing Works. *Pages 581–590 of: CHI '06 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

DuPaul, Neil. 2012. Common malware types: Cybersecurity 101. Online. `https://www.veracode.com/blog/2012/10/common-malware-types-cybersecurity-101` (accessed on 2017-10-01).

Elledge, Anthony. 2004. *Phishing: An Analysis of a Growing Threat*. Technical report. SANS Infosec Reading Room.

Fan, Jianqing, Han, Fang, & Liu, Han. 2014. Challenges of Big Data analysis. *National Science Review*, **1**(2), 293–314.

Fayyad, Usama, Piatetsky-Shapiro, Gregory, & Smyth, Padhraic. 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, **17**(3), 37.

Ferguson, Aaron. 2005. Fostering E-Mail Security Awareness: The West Point Carronade. *Educause Quarterly*, **1**, 1–4.

Frost & Sullivan. 2016. *Advanced Malware Sandbox Market Analysis*. Technical report. Frost & Sullivan.

Goebel, Michael, & Gruenwald, Le. 1999. A Survey of Data Mining and Knowledge Discovery Software Tools. *SIGKDD Explor. Newsl.*, **1**(1), 20–33.

Gohstand, Jonathan. 2010. Getting the most from Active Directory in the enterprise. Online. `https://esj.com/articles/2010/06/15/active-directory-in-the-enterprise.aspx` (accessed on 2017-06-06).

Goodman, Joshua, Cormack, Gordon V, & Heckerman, David. 2007. Spam and the Ongoing Battle for the Inbox. *Magazine Communications of the ACM*, **50**(2), 24–33.

Graziano, Mariano, Canali, Davide, Bilge, Leyla, Lanzi, Andrea, & Balzarotti, Davide. 2015. Needles in a Haystack: Mining Information from Public Dynamic Analysis Sandboxes for Malware Intelligence. *Pages 1057–1072 of: Proceedings of the 24th USENIX Security Symposium (USENIX '15)*. Usenix Association.

Greaux, Scott. 2013. Use metrics to measure and improve security awareness. Online. `https://phishme.com/use-metrics-measure-improve-effectiveness-security-awareness/` (accessed on 2017-09-20).

Griffin, Nicholas. 2016. Locky distributor uses newly released Quant loader sold on Russian underground. Online. https://blogs.forcepoint.com/security-labs/locky-distributor-uses-newly-released-quant-loader-sold-russian-underground (accessed on 2017-09-11).

Group IB. 2016. *BUHTRAP - The evolution of targeted attacks against financial institutions*. Technical report. Group IB.

Group IB, & Fox-IT. 2014. *Anunak: APT against financial institutions*. Technical report. Group IB & Fox-IT.

Grunzweig, Josh. 2016. Investigating the LuminosityLink remote access trojan configuration. Online. https://researchcenter.paloaltonetworks.com/2016/07/unit42-investigating-the-luminositylink-remote-access-trojan-configuration/ (accessed on 2017-09-20).

Gumbs, Gabriel. 2017. LDAP monitoring for security. Online. https://blog.stealthbits.com/LDAP-monitoring-for-security/ (accessed on 2017-06-06).

Hardikar, Aman. 2008. *Malware 101 - Viruses*. Technical report. SANS Institute Infosec Reading Room.

Hawthorne, Andrew. 2016. Start seeing the threats before they hit you. *Pages 1–11 of: IBM i2 User Group Conference*.

Hewlett Packard Enterprise. 2016. *Comparing SIEM, big data, and behavior analytics security management solutions*. Technical report. Hewlett Packard Enterprise.

Hinchliffe, Alex. 2017. EMEA Bi-Monthly Threat Reports: Turkey, Saudi Arabia & United Arab Emirates. Online. https://researchcenter.paloaltonetworks.com/2017/07/unit42-emea-bi-monthly-threat-reports-turkey-saudi-arabia-united-arab-emirates/ (accessed on 2017-09-09).

Hoppe, Tobias, Pastwa, Alexander, & Sowa, Sebastian. 2009. Business Intelligence Based Malware Log Data Analysis as an Instrument for Security Information and Event Management. *Journal On Advances in Security*, **2**(2), 203–213.

Hutchins, Eric, Cloppert, Michael, & Amin, Rohan. 2011. *Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains*. Technical report. July 2005. Lockheed Martin.

Informatica. 2013. *Data Fusion for Cyber Intelligence*. Technical report. Informatica.

Irinco, Bernadette. 2015. Carbanak targeted attack campaign hits banks and financial institutions. Online. `https://www.trendmicro.com/vinfo/us/threat-encyclopedia/web-attack/3142/carbanak-targeted-attack-campaign-hits-banks-and-financial-institutions` (accessed on 2017-04-10).

Jagatic, Tom N, Johnson, Nathaniel A, Jakobsson, Markus, & Menczer, Filippo. 2007. Social Phishing. *Commun. ACM*, **50**(10), 94–100.

Jiang, Genwei, Mohandas, Rahul, Leathery, Jonathan, Berry, Alex, & Galang, Lennard. 2017. CVE-2017-0199: In the wild attacks leveraging HTA handler. Online. `https://www.fireeye.com/blog/threat-research/2017/04/cve-2017-0199-hta-handler.html` (accessed on 2017-09-11).

Karnouskos, Stamatis. 2011. Stuxnet worm impact on industrial cyber-physical system security. *IECON Proceedings (Industrial Electronics Conference)*, 4490–4494.

Kaspersky Lab. 2015a. *Carbanak Apt the Great Bank Robbery*. Technical report. Kaspersky Lab.

Kaspersky Lab. 2015b. *Duqu 2.0 a sophisticated cyberespionage actor returns*. Technical report. Kaspersky Lab.

Kaspersky Lab. 2016. *The amount of malicious spam reaches all time high in Q3, 2016*. Technical report. Kaspersky Lab.

Kaspersky Lab. 2017a. *Lazarus under the hood*. Technical report. Kaspersky Lab.

Kaspersky Lab. 2017b. Spam and phishing in Q1 2017. Online. `https://securelist.com/spam-and-phishing-in-q1-2017/78221/` (accessed on 14/08/2017).

Kendall, Chris, & McMillan, Chad. 2007. Practical Malware Analysis. *Pages 1–65 of: Blackhat USA 2007*.

Kessem, Limor. 2017a. All in a spammer's workweek: Where do the busiest spammers work around the clock? Online. `https://securityintelligence.com/all-in-a-spammers-workweek-where-do-the-busiest-spammers-work-around-the-clock/` (accessed on 2017-08-23).

Kessem, Limor. 2017b. TrickBot: Trojan widens its attack scope in Spain, brings redirection attacks to local banks. Online. `https://securityintelligence.com/trickbot-habla-espanol-trojan-widens-its-attack-scope-in-spain-brings-redirection-attacks-to-local-banks/` (accessed on 2017-09-12).

Kharpal, Arjun. 2015. Hackers being hunted after stealing $30.7M via malware. Online. `http://www.cnbc.com/2015/10/14/hackers-being-hunted-after-using-dridex-malware-to-steal-over-30m.html` (accessed on 10/04/2017).

Khonji, Mahmoud, Iraqi, Youssef, & Jones, Andrew. 2013. Phishing Detection: A Literature Survey. *IEEE Communications Surveys and Tutorials*, **15**(4), 2091–2121.

Kirk, Andy. 2016. Exploratory data analysis: Widening your view point. Online. `http://www.statisticsviews.com/details/feature/9971321/Exploratory-Data-Analysis-Widening-Your-View-Point.html` (accessed on 2017-09-01).

Krebs, Brian. 2013. Polish takedown targets "virut" botnet. Online. `https://krebsonsecurity.com/2013/01/polish-takedown-targets-virut-botnet/` (accessed on 2017-09-15).

Le Blond, Stevens, Uritesc, Adina, Gilbert, Cedric, Leong Chua, Zheng, Saxena, Prateek, & Kirda, Engin. 2014. A Look at Targeted Attacks Through the Lense of an NGO. *Pages 543–558 of: Proceedings of the 23rd USENIX Security Symposium (USENIX '14)*. Usenix Association.

Lee, Martin, & Lewis, Daren. 2011. Clustering Disparate Attacks: Mapping the Activities of the Advanced Persistent Threat. *Pages 1–22 of: Virus Bulletin Conference*.

Lee, Robert M. 2012. *The History of Stuxnet: Key Takeaway for Cyber Decision Makers*. Technical report. Armed Forces Communications and Electronics Association.

Leyden, John. 2004. Microsoft attack worm rides on the back of MyDoom. Online. `https://www.theregister.co.uk/2004/07/28/ms_worm_uses_mydoom/` (accessed on 15/09/2017).

Leyden, John. 2016. Inside Adwind: A DIY malware toolkit used by 1,800 crooks to spy on 443k victims. Online. `https://www.theregister.co.uk/2016/02/08/adwind/` (accessed on 12/09/2017).

Llascu, Ionut. 2015. Leaked full version of NanoCore RAT used to target energy companies. Online. http://news.softpedia.com/news/Leaked-Full-Version-of-NanoCore-RAT-Used-to-Target-Energy-Companies-476606.shtml (accessed on 2017-09-11).

Lumeta. 2015. *Continuous Cyber Situational Awareness*. Technical report. Lumeta.

Magnúsardóttir, Arna. 2017. Locky 2? Jaff ransomware launched from Necurs botnet. Online. https://blog.cyren.com/articles/locky-2-jaff-ransomware-launched-from-necurs-botnet (accessed on 2017-09-23).

McAfee Labs. 2016. *Understanding Ransomware and Strategies to Defeat it*. Technical report. McAfee Labs.

McDonald, Geoff, O Murchu, Liam, Doherty, Stephen, & Chien, Eric. 2013. *Stuxnet 0.5: The Missing Link*. Technical report. Symantec Security Response.

Meidam, Katrien. 2015. *Phishing as a Service: Designing an ethical way of mimicking targeted phishing attacks to train employees*. Master Thesis, Delft University of Technology.

Merrit, Eric. 2017. Latest malware uses compiled AutoIT script to masquerade as Photoshop CS6 installer. Online. https://www.carbonblack.com/2017/04/05/latest-malware-uses-compiled-autoit-script-masquerade-photoshop-cs6-installer/ (accessed on 2017-09-11).

Microsoft. 2000. Active Directory. Online. https://https//msdn.microsoft.com/en-us/library/bb742424.aspx (accessed on 06/06/2017).

Microsoft. 2016. Schema attributes. Online. https://msdn.microsoft.com/en-us/library/ms675090(v=vs.85).aspx (accessed on 2016-08-12).

MITRE. 2017a. DLL injection. Online. https://attack.mitre.org/wiki/Technique/T1055 (accessed on 2017-09-09).

MITRE. 2017b. Process hollowing. Online. https://attack.mitre.org/wiki/Technique/T1093 (accessed on 2017-09-11).

Munro, Jay. 2004. MyDoom.A: Fastest spreading virus in history. Online. https://www.pcmag.com/article2/0,2817,1485719,00.asp (accessed on 2017-09-14).

Nettitude. 2016. *SWIFT Threat Advisory Report*. Technical report. Nettitude.

NIST. 2017. EDA handbook. Online. http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm (accessed on 2017-08-15).

Oprea, Alina, Li, Zhou, Yen, Ting Fang, Chin, Sang H., & Alrwais, Sumayah. 2015. Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data. *Pages 45–56 of: Proceedings of the 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*.

PCI Security Standards Council LLC. 2016. *Effective Daily Log Monitoring Guidance*. Technical report. PCI Security Standards Council LLC.

Perlroth, Nicole, & Shane, Scott. 2017. How Israel caught Russian hackers scouring the world for U.S. secrets. Online. https://mobile.nytimes.com/2017/10/10/technology/kaspersky-lab-israel-russia-hacking.html (accessed on 2017-10-11).

Phishlabs. 2016. *Hacking the Human*. Technical report. Phishlabs.

Ponemon Institute LLC. 2013. *Big Data Analytics in Cyber Defense*. Technical report. Ponemon Institute LLC.

Proofpoint. 2015. *The Human Factor*. Technical report. Proofpoint.

Proofpoint. 2017. New ransomware targeting education and healthcare verticals. Online. https://www.proofpoint.com/us/threat-insight/post/defray-new-ransomware-targeting-education-and-healthcare-verticals (accessed on 2017-09-21).

Ratanaworabhan, Paruj, Livshits, Benjamin, & Zorn, Benjamin. 2009. NOZZLE: A Defense Against Heap-spraying Code Injection Attacks. *Pages 169–186 of: Proceedings of the 18th Conference on USENIX Security Symposium*. SSYM'09. Berkeley, CA, USA: USENIX Association.

Rattray, Greg. 2014. Building an Effective Corporate Cyber Threat Intelligence Practice. *Pages 1–23 of: SANS Cyber Threat Intelligence Summit*.

Rekouche, Koceilah. 2011. Early phishing. Online. https://arxiv.org/ftp/arxiv/papers/1106/1106.4692.pdf (accessed on 2017-04-21).

Rieck, Konrad, Holz, Thorsten, Willems, Carsten, Düssel, Patrick, & Laskov, Pavel. 2008. Learning and Classification of Malware Behavior. *Pages 108–125 of: Proceedings of the 5th International Conference on Detection of Intrusions and Mal-*

ware, and Vulnerability Assessment. DIMVA '08. Berlin, Heidelberg: Springer-Verlag.

Robinson, Rick. 2015. Malicious attachments make a comeback as top attack vector. Online. `https://securityintelligence.com/malicious-attachments-make-a-comeback-as-top-attack-vector/` (accessed on 2017-04-06).

RSA. 2017. *The Carbanak/Fin7 Syndicate A Historical Overview of an Evolving Threat*. Technical report. RSA.

Russinovich, Mark. 2016. Tracking hackers on your network with sysinternals sysmon. *Pages 1–39 of: RSA Conference 2016*.

Securosis. 2015. *Applied Threat Intelligence*. Technical report. Secrosis.

Shackleford, Dave. 2014. *Analytics and Intelligence Survey*. Technical report. SANS Institute Infosec Reading Room.

Shackleford, Dave. 2015. *Who's Using Cyberthreat Intelligence and How?* Technical report. SANS Institute Infosec Reading Room.

Shackleford, Dave. 2016. *The SANS state of Cyber Threat Intelligence Survey: CTI Important and Maturing*. Technical report. SANS Institute Infosec Reading Room.

Shackleford, Dave. 2017. *Cyber Threat Inteligence Uses, Successes and Failures: The SANS 2017 CTI Survey*. Technical report. SANS Institute Infosec Reading Room.

Shaw, R S, Chen, Charlie C, Harris, Albert L, & Huang, Hui-Jou. 2009. The Impact of Information Richness on Information Security Awareness Training Effectiveness. *ELSEVIER Computers & Education*, **52**(1), 92–100.

Solutionary. 2005. *How Malware Analysis Benefits Incident Response*. Technical report. Solutionary.

Splunk. 2017. Monitor privileged accounts for suspicious activity. Online. `https://docs.splunk.com/Documentation/ES/4.7.1/Usecases/PrivilegedUsers` (accessed on 06/06/2017).

Stone-Gross, Brett, Holz, Thorsten, Stringhini, Gianluca, & Vigna, Giovanni. 2011. The Underground Economy of Spam: A Botmaster's Perspective of Coordinating

Large-scale Spam Campaigns. *Page 4 of: Proceedings of the 4th USENIX Conference on Large-scale Exploits and Emergent Threats*. LEET'11. Berkeley, CA, USA: USENIX Association.

Svajcer, Vanja. 2017. Modified Zyklon and plugins from India. Online. `http://blog.talosintelligence.com/2017/05/modified-zyklon-and-plugins-from-india.html` (accessed on 12/09/2017).

Symantec. 2007. W32.Virut. Online. `https://www.symantec.com/security_response/writeup.jsp?docid=2007-041117-2623-99` (accessed on 15/09/2017).

Symantec. 2014. *Attack on point of sales systems*. Technical report. Symantec.

Symantec. 2015. *Symantec Intelligence Report*. Technical report. Symantec.

Symantec. 2016a. *Internet Security Threat Report: Ransomware and Businesses*. Technical report. Symantec.

Symantec. 2016b. *Internet Security Threat Report Volume 21*. Technical report. Symantec.

Symantec. 2016c. *The increased use of PowerShell in attacks*. Technical report. Symantec.

Symantec. 2017. *Internet Security Threat Report Volume 22*. Technical report. Symantec.

Symantec Security Response. 2016. Odinaff: New trojan used in high level financial attacks. Online. `https://www.symantec.com/connect/blogs/odinaff-new-trojan-used-high-level-financial-attacks` (accessed on 2017-04-10).

Szappanos, Gábor. 2017. *CVE-2017-0199: Life of an exploit*. Technical report. Sophos.

Tableau. 2017a. Connecting Excel CSV and text files. Online. `https://www.tableau.com/learn/tutorials/on-demand/connecting-excel-csv-and-text-files` (accessed on 2017-08-15).

Tableau. 2017b. Join your data. Online. `https://onlinehelp.tableau.com/current/pro/desktop/en-us/joining_tables.html` (accessed on 2017-08-15).

Talos Group. 2015. Malware meets sysAdmin – Automation tools gone bad. Online. `http://blogs.cisco.com/security/talos/sysadmin-phish` (accessed on 2017-09-29).

Tarala, James. 2011. *A Real-Time Approach to Continuous Monitoring*. Technical report. SANS Institute InfoSec Reading Room.

Telus Security Labs. 2016. Trojan.JS.Cryxos.A. Online. `http://telussecuritylabs.com/threats/show/TSL20170726-02` (accessed on 2017-09-12).

Teymourlouei, Haydar, & Jackson, Lethia. 2017. How Big Data Can Improve Cyber Security. *Pages 9–13 of: International Conference on Advances in Big Data Analytics*.

Trend Micro. 2015. *The Cuckoo Miner Campaign - Nigerian Cyber Criminals Targeting Banks*. Technical report. Trend Micro Labs.

Trend Micro. 2017. Advanced threat detection by deep discovery. Online. `https://www.trendmicro.com/en_us/business/products/network/deep-discovery.html` (accessed on 2017-04-11).

Vanier, Sarah. 2016. The most devastating cyber attacks on banks. Online. `https://sentinelone.com/blogs/the-most-devastating-cyber-attacks-on-banks/` (accessed on 2017-04-06).

Varga, Margaret, Winkelholz, Carsten, & Träber-Burdin, Susan. 2016. *Cyber Situation Awareness*. Technical report. NATO Science & Technology Organization.

Villeneuve, Nart. 2011. *Trends in Targeted Attacks*. Technical report. Trend Micro.

Wakelin, Chris. 2016. Locky ransomware is becoming more sophisticated. Online. `https://www.proofpoint.com/us/threat-insight/post/Locky-Ransomware-Cybercriminals-Introduce-New-RockLoader-Malware` (accessed on 2017-09-12).

Walker, Christopher. 2016. *Threat Intelligence: Planning and Direction*. Technical report. SANS Institute Infosec Reading room.

Webroot. 2013. *Threat Intelligence: What is it, and How Can it Protect You from Today's Advanced Cyber-Attacks?* Technical report. Webroot.

Wu, Rubio, & Chen, Marshall. 2017. Spam campaign delivers cross-platform remote access Trojan Adwind. Online. `http://blog.trendmicro.com/trendlabs-security-intelligence/spam-remote-access-trojan-adwind-jrat/` (accessed on 2017-09-12).

Wueest, Candid. 2015. *The State of Financial Trojans*. Technical report. Symantec.

Wueest, Candid. 2016. *Financial Threats 2015*. Technical report. Symantec.

Zaharia, Andra. 2016. Atmos carries on the ZeuS legacy. Online. `https://heimdalsecurity.com/blog/security-alert-citadel-trojan-resurfaces-atmos-zeus-legacy/` (accessed on 2017-09-12).

Zamora, Wendy. 2017. What are exploits? (And why you should care). Online. `https://blog.malwarebytes.com/101/2017/03/what-are-exploits-and-why-you-should-care/` (accessed on 29/09/2017).

Zhang, Xiaopeng, & Liu, Hua. 2017. New Loki variant being spread via PDF file. Online. `https://blog.fortinet.com/2017/05/17/new-loki-variant-being-spread-via-pdf-file` (accessed on 2017-09-12).

Appendix A

SHA256 Cryptographic Identifiers

The SHA256 hashes used to conduct this research are available on Github at the following link:

`https://github.com/JapieVermeulen/MSc-2017-Malware-Hashes`

Table A.1: Appendix SHA256 filename checksums

Filename	File checksums
appendix_256hashes.csv	706074A7EE1C774CF7C6F43BDF68FE7DD56693C186FE4C13C16669DAC467E528
appendix_256hashes.pdf	DA6032063001179E0DB46135A723CEBA5DA9BB91FF72B49BD303FE20EA120675