A False Positive Prevention Framework for Non-Heuristic Anti-Virus Signatures

Symantec, A Case Study

Thomas Parsons
Acknowledgements

Thanks to my wife Eithne for all her support and patience over the past two years.

I would also like to thank my supervisor, Hugh Gibbons, and also Frank Bannister and Dan Remenyi for their advice and guidance.

Also thanks to the many people at Symantec for their time and effort in helping with this research particularly those who participated in the interviews.

Abstract

False positives, the erroneous detection of clean files, have been referred to as the Achilles heel of the anti-virus industry. Some believe the problem false positives represent is growing. This belief is likely underpinned by the growth in anti-virus signatures, due to the exponential growth in malicious code over the past 2 years, and the corresponding impact this invariably has on false positives. False positives can have a serious impact on users (system downtime, data loss) and on the anti-virus vendor responsible (damage to brand).

This research attempts to identify the root cause of false positives from non-heuristic anti-virus signatures. Non-heuristic signatures are characterised as being re-active and written in response to a known threat. These signatures are also referred to as fingerprints or pattern files. Non-heuristic technology itself is the most pervasive technology used by anti-virus vendors since the industry’s inception over 20 years ago.

Using the available literature and root cause data from a secondary data source at Symantec, in tandem with qualitative data from interviews, the research looks at developing a framework for preventing false positives from non-heuristic signatures. A case study was used to investigate and collect data.

The context of this study is specific to Symantec. To determine the root causes data will be leveraged from a key system in Symantec called the False Positive Logging System (FPLS). Other qualitative data will be solicited through semi-structured interviews with domain experts within Symantec’s Security Technology and Response (STAR) organisation.

This study shows that legacy solutions to address false positives at Symantec were traditionally aimed at the ‘detection’ of false positives. However solutions based around the ‘prevention’ of false positives are more efficient. These solutions trigger far earlier in the signature generation lifecycle. Most importantly defect prevention directly targets the leading cause of false positives as identified by this study. A defect prevention approach is also supported by previous work and standards such as the Defect Prevention Process (DPP) and the Capability Maturity Model Integration (CMMI). In essence the research proposes structural change of signature generation processes at Symantec.
## Contents

Acknowledgements .............................................................................................................. 2  
Abstract ................................................................................................................................... 2  

List of Figures ........................................................................................................................... 4  
List of Tables ............................................................................................................................ 5  
List of Acronyms ....................................................................................................................... 6  

1. Introduction .......................................................................................................................... 7  
   1.1. Background ...................................................................................................................... 7  
   1.2. The problem ................................................................................................................... 7  
   1.3. The research questions/objectives .................................................................................. 7  
   1.4. Why and to whom is it important? .................................................................................. 8  
   1.5. Scope of the study .......................................................................................................... 8  
   1.6. Context of study and author/organisation ...................................................................... 8  
   1.7. Timeframe of the research ............................................................................................ 9  
   1.8. Roadmap of chapters .................................................................................................... 9  

2. Literature Review ................................................................................................................ 10  
   2.1. False positive in the blacklisting paradigm .................................................................. 10  
   2.2. Non-heuristic technology ............................................................................................ 11  
   2.3. False positives in related technologies ........................................................................ 11  
   2.4. Non-primary consideration of false positives in anti-virus research ............................ 11  
   2.5. Signature generation processes ................................................................................... 12  
   2.6. Clean data ..................................................................................................................... 13  
   2.7. Defect Prevention Process (DPP) ................................................................................ 15  
   2.8. DPP adoption in CMMI ............................................................................................... 16  
   2.9. Defect cost relative to identification point in process .................................................. 17  
   2.10. Summary .................................................................................................................... 17  

3. Methodology ...................................................................................................................... 19  
   3.1. Research question and objectives ................................................................................. 19  
   3.2. Approaches considered ............................................................................................... 19  
   3.3. Case study ................................................................................................................... 19  
   3.4. Practitioner-researcher role .......................................................................................... 20  
   3.5. Data types and analysis ............................................................................................... 20  
   3.6. Limitations .................................................................................................................. 21  
   3.7. Ethical considerations .................................................................................................. 21  

4. Field Research .................................................................................................................. 22  
   4.1. Secondary data ............................................................................................................. 22  
   4.2. Semi-structured interviews .......................................................................................... 24  

5. Findings and Analysis ........................................................................................................ 28  
   5.1. Root causes for “high severity” false positives ............................................................... 28  
   5.2. Root causes for all false positives ............................................................................... 28  
   5.3. Theme 1: Historical context of false positives .............................................................. 29  
   5.4. Theme 2: Cost of a FP ................................................................................................. 31  
   5.5. Theme 3: Legacy investment in FP avoidance .............................................................. 32  
   5.6. Theme 4: Root causes ................................................................................................. 32  
   5.7. Theme 5: Ideal approaches to FP avoidance ............................................................... 33  
   5.8. Theme 6: Future FP trends and risks .......................................................................... 33  
   5.9. Summary ...................................................................................................................... 34  

6. Conclusions and Future Work .......................................................................................... 36  
   6.1. Framework development ............................................................................................. 36  
   6.2. Framework outline ....................................................................................................... 37  
   6.3. Framework implementation at Symantec .................................................................... 38  
   6.4. Generalisability ........................................................................................................... 39  
   6.5. Advancement of current thinking on topic ................................................................... 39  
   6.6. Limitations .................................................................................................................. 40  
   6.7. Future directions of research ....................................................................................... 40  

References .................................................................................................................................. 41  
Appendix .................................................................................................................................... 45
List of Figures

| Figure 1: Conditional testing: Blacklisting and whitelisting approaches | 10 |
| Figure 2: High level view of the signature generation process | 12 |
| Figure 3: Root causes of high severity FPs. | 28 |
| Figure 4: Root causes of all false positives. | 29 |
| Figure 5: Representation of main findings and analysis. | 34 |
| Figure 6: Framework outline for preventing false positives | 37 |
| Figure 7: Severity level of false positives. | 39 |
List of Tables

Table 1: Review topics and associated research field .......................................................... 10
Table 2: Representative average relative defect-removal costs ........................................... 17
Table 3: Uses of different types of interview ........................................................................ 20
Table 4: Examples of FPLS root causes ........................................................................ 23
Table 5: FPLS severity levels and descriptions .................................................................... 23
Table 6: Interview theme development ............................................................................. 25
Table 7: Interview participants ......................................................................................... 25
Table 8: Costs of a false positive ....................................................................................... 31
Table 9: Legacy investment in FP avoidance and its deficiencies ....................................... 32
Table 10: Components for the ideal approach to false positive avoidance ............................ 33
## List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>Anti-spam</td>
</tr>
<tr>
<td>CMMI</td>
<td>Capability Maturity Model Integration</td>
</tr>
<tr>
<td>DPP</td>
<td>Defect Prevention Process</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>FPLS</td>
<td>False Positive Logging System</td>
</tr>
<tr>
<td>IBM</td>
<td>International Business Machines</td>
</tr>
<tr>
<td>IDC</td>
<td>International Data Corporation</td>
</tr>
<tr>
<td>IDS</td>
<td>Intrusion Detection System</td>
</tr>
<tr>
<td>IPS</td>
<td>Intrusion Prevention System</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organisation for Standardisation</td>
</tr>
<tr>
<td>ISV</td>
<td>Independent Software Vendor</td>
</tr>
<tr>
<td>MB</td>
<td>Megabyte</td>
</tr>
<tr>
<td>OS</td>
<td>Operating System</td>
</tr>
<tr>
<td>SEI</td>
<td>Software Engineering Institute</td>
</tr>
<tr>
<td>SP</td>
<td>Software Process</td>
</tr>
<tr>
<td>SPI</td>
<td>Software Process Improvement</td>
</tr>
<tr>
<td>SRE</td>
<td>Security Response Engineer</td>
</tr>
<tr>
<td>SRM</td>
<td>Security Response Manager</td>
</tr>
<tr>
<td>STAR</td>
<td>Security Technology and Response</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Background

False positives, the erroneous detection of clean files, are an important quality concern for the anti-virus industry and its customers. The implications of releasing an anti-virus signature that causes a false positive (FP) condition can be severe. Examples of the impact from the customer perspective are system downtime, data loss, and lack of trust in the software that is designed to protect their information. From the anti-virus vendor perspective examples include loss of business, negative publicity, and perhaps litigation (Yang, 2007).

Anti-virus vendors have traditionally adopted a re-active approach to malicious code based on a blacklisting model. This means that signatures are typically developed in response to a threat. Over the past decade the industry has moved towards a more pro-active approach to detecting malicious code. This approach utilises a combination of advanced heuristic (Arnold and Tesauro, 2000) and behavioural based technologies (Williamson, 2005) in addition to non-heuristic signatures (Perry, 2008).

The year 2007 witnessed exponential growth in malicious code in comparison to the previous twenty years (Bradley, 2008). There has been a similar growth in non-heuristic signatures developed by anti-virus vendors in an attempt to keep customers protected.

Non-heuristic signatures are characterised as being re-active and written in response to a known threat. These signatures are also referred to as fingerprints or pattern files. They have been described as “the most cost-effective way for a manufacturer to provide detection” and “the mainstay and staple of malware detection for 20 years” (Perry, 2008). Continuously increasing the volume of malicious code is essentially a tactic used by malicious code authors to stretch and test the capabilities of anti-virus vendors. The goal of this tactic is to get malicious code installed on a target system so the resources on or associated with that system can be exploited (Yan et al., 2008). Anti-virus vendors have been churning out anti-virus signatures at such a high rate in response to these threats that the blacklisting paradigm adopted by the industry since its inception has been called into question in terms of scalability (Uscilowski et al., 2008) (Nachenberg et al., 2009).

With the growth in malicious code and the corresponding release of high volumes of signatures by anti-virus vendors there is a growing concern in relation to number of false positives that may occur (Yan et al., 2008). While anti-virus technology is generally less prone to false positives than say intrusion prevention or anti-spam technologies, any growth in anti-virus false positives is a major concern. The reason being is the high market penetration and install base of anti-virus technology. For instance Symantec endpoint security solutions are installed on over 120 million client, server and gateway systems worldwide (Symantec, 2008). In addition anti-virus technology has traditionally been more visible when compared with other related technologies as it is typically deployed at many levels within an organisation such as desktop, fileserver, and gateway. This work focuses on the prevention of false positives from non-heuristic signatures.

1.2. The problem

Anti-virus technology is designed to detect malicious files. When such a condition exists it is referred to as a true positive (TP). False positives on the other hand can be defined as the erroneous detection of an innocuous file. Though higher severity false positives generate much attention, it is likely that vendors within the anti-virus industry release far higher numbers of lower severity false positives (see section 4.1.5 for Symantec’s severity level classification scheme for false positives). It is largely accepted across the industry that false positives are inevitable. However the problem of how to vastly reduce or even eliminate higher severity false positives and how to minimise lower severity false positives from non-heuristic signatures is a key question that this research will attempt to answer.

1.3. The research questions/objectives

The main goal of this research is to establish a framework that leverages software process improvement (SPI) techniques to eliminate high severity false positives from non-heuristic signatures released by Symantec. In addition this should be achieved in a cost efficient manner while also having a positive effect on the overall number of...
false positives from non-heuristic signatures. The framework should also provide visibility to management and engineers of signature quality throughout the content generation lifecycle.

Four main inputs influenced the development of the framework. The first of these is in-field data from Symantec on the root causes of false positives. The second input is literature on quality assurance models and standards that lend themselves to such a framework such as defect prevention. The third input is experience that the author has in this area over the past three years. The final input is information from knowledge leaders and others within Symantec in relation to false positives, their causes and prevention.

1.4. Why and to whom is it important?

Many anti-virus vendors use non-heuristic signatures as an important part of their approach to tackling malicious code. While the context of the research is specific to Symantec it is likely to be applicable for all anti-virus vendors using this technology. False positives are costly to anti-virus vendors in many different ways, as this research will show, and can also cause serious problems for users of endpoint security products.

The research borrows elements, like root cause analysis and defect prevention, from industry standard software process improvement (SPI) approaches such as the Capability Maturity Model Integration (CMMI) which are used in different fields across the software industry. Thus this research may also be of interest to quality assurance practitioners or advocates of such models and approaches from these fields.

The most interested party in this research though is likely to be Symantec itself. While non-heuristic technology is a mature technology and much research has been published on it by Symantec and others, little research has been conducted specifically on false positives as a result of using this technology. This research may prompt further work on false positives related to non-heuristic signatures, or other anti-virus technologies, by quality assurance practitioners or others within the industry.

1.5. Scope of the study

The scope of this study limits itself to false positives from non-heuristic signatures. There are a number of other anti-virus technologies (e.g. heuristic, behavioural based etc.) used in endpoint security products today that generate false positives but these are not within the scope of this research (Perry, 2008). Likewise false positives are not confined to Symantec and are an industry wide problem. This work is the first detailed case study in this area and is company (Symantec) focused. As discussed in the Conclusions and Future Work chapter an obvious next step is to undertake a broader pan-industry study.

From a paradigmatic perspective the scope of the research is limited to false positives as a result of the blacklisting approach. This approach has been used since the anti-virus industry was established over 20 years ago. While claims that the blacklisting model is broken have been argued (Perry, 2008), there is no doubt that some anti-virus vendors are looking beyond this traditional approach (Nachenberg et al., 2009). Symantec, for one, is adopting a reputation based approach to help protect customers pro-actively against unknown threats (Nachenberg et al., 2009). This approach does not fit with the blacklisting approach and effectively represents a new paradigm. A new industry has developed on the back of recent discussions on the efficacy of the blacklisting paradigm i.e. the whitelisting industry. Whitelisting is the flip side of the blacklisting coin (as figure 1 in section 2.1 demonstrates). It’s based on the knowledge of all valid or clean files only and any file it does not have knowledge of is flagged as bad and blocked by the whitelisting product (Bloor, 2006). A move to a whitelisting approach therefore represents a shift rather than a new paradigm as is the case with the reputation based approach. Whitelisting is prone to false positives too. However a false positive in the context of this approach is to refer to a malicious file as good. False positives within the context of whitelisting are not within the scope of this research.

1.6. Context of study and author/organisation

The author is an employee of Symantec. Symantec is a large multi-national company employing 17,500 people across 40 countries and generated revenues of $6.2 billion in the company’s accounting year for 2009. The company is focused on providing security, storage, and systems management solutions to help businesses and
consumers secure and manage their information (Symantec, 2009a).

In relation to the security aspect of the Symantec’s business it competes in both the consumer and enterprise information security markets. Symantec is ranked number one in market share by IDC (International Data Corporation) in eight security markets including Worldwide Endpoint Security Software, Worldwide Corporate Endpoint Security Software and also Worldwide Consumer Endpoint Security Software (Symantec, 2009b).

Anti-virus technology is a key offering within Symantec’s leading security suites; Norton Internet Security (consumer focused) and Symantec Endpoint Protection (enterprise focused). The latter is classified as a ‘leader’ in Gartner’s Magic Quadrant Dec. 2007 (Firstbrook et al., 2007). Given the importance of anti-virus technology to their key security product offerings, Symantec takes any issues with quality related to these very seriously.

In the latter half of 2007 Symantec began to organise itself to better address false positives in response to a serious false positive its customers experienced in China. This false positive, caused by a non-heuristic signature and released in May 2007, resulted in negative press worldwide for Symantec. In China the state-sponsored Xinhua News Agency commented that the false positive was “a heavy blow to people's daily work and ongoing business”(Keizer, 2007b). This was Symantec’s most significant false positive since early 2000. As a response to the Chinese false positive Symantec’s Security Response organisation established a “Quality Initiative Strategy” that was presented to and approved by Symantec’s then CEO, John W. Thompson, in September 2007. This initiative had three main elements, the principal one being exclusively focused around false positives.

A company re-organisation in 2008 saw the Security Response group move from Services to the office of the CTO and was merged with Symantec’s former Security Technology Group (STG) to form a new group called Security Technology and Response (STAR). This group is responsible for the development of all engine and content for Symantec’s endpoint security offerings in the consumer and enterprise markets. As a result of this re-organisation the “Quality Initiative” team was effectively replaced by a new “False Positive Focus Team” where the focus was narrower but more precise. The author is a member of this team and was a member of the former.

1.7. Timeframe of the research

In the context of this dissertation, the effective window for research ran from October 2008 to August 2009. However efforts at Symantec in relation to the research problem extend beyond this period.

1.8. Roadmap of chapters

Chapter 2 Literature Review discusses previous work on false positives. False positives are referenced in many papers related to anti-virus technology although are rarely the primary focus of this literature. That aside the author will review, critique and discuss the key ideas from this literature.

Chapter 3 Methodology briefly outlines the approaches considered for the research. The approach ultimately selected is discussed and justified.

Chapter 4 Field Research outlines the specific research tactics used. The use of company data and also interviews with leading thinkers within Symantec as the main data collection devices is discussed.

Chapter 5 Findings and Analysis reports the key outcomes of the dissertation. These outcomes include the main root causes of false positives from non-heuristic anti-virus signatures. Based on these results the key mechanisms that can prevent false positives from occurring are discussed along with their injection points in the anti-virus signature lifecycle. These mechanisms will then form the basis of the proposed false positive prevention framework.

Chapter 6 Conclusion and Future Work completes the dissertation by demonstrating the research question has been answered, outlining the limitations of the research, and by identifying a number of interesting areas for further research.
2. Literature Review

This study straddles both anti-virus and quality assurance fields. In line with this, the review covers literature from both these fields. The key topics reviewed in this chapter and their associated field are presented in the table below.

<table>
<thead>
<tr>
<th>Research field</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-virus</td>
<td>False positives and their place in the current anti-virus paradigm.</td>
</tr>
<tr>
<td>Anti-virus</td>
<td>Technology (anti-virus and related) prone to false positives.</td>
</tr>
<tr>
<td>Anti-virus</td>
<td>Processes (human and automated) at the root of false positive generation.</td>
</tr>
<tr>
<td>Anti-virus</td>
<td>Clean data and its importance to avoiding false positives.</td>
</tr>
<tr>
<td>Quality Assurance/SPI</td>
<td>Where DPP fits today in software process improvement models (e.g. CMMI).</td>
</tr>
<tr>
<td>Quality Assurance/SPI</td>
<td>Cost of a removing a defect.</td>
</tr>
</tbody>
</table>

2.1. False positive in the blacklisting paradigm

In the antivirus industry today a false positive is considered to be the erroneous detection of a clean or innocuous file (or files). The terminology may differ from time to time but this meaning is broadly accepted and understood to the point where a definition isn’t often stated in literature that discusses false positives. Where a definition is attempted it tends to be brief (David, 1996; Chess and White, 2000) (Bureau and Harley, 2008) but in some cases is more clearly defined (Schultz et al., 2001). The term ‘false positive’ is used in some literature (Preda et al., 2008), called ‘false alarms’ in other literature (David, 1996), and also may be used interchangeably (Venugopal et al., 2006).

With the blacklisting approach adopted by the anti-virus industry since its inception, a false positive condition can only exist when a clean file is present and typically where a signature has been created by an anti-virus vendor that detects or convicts this clean file (or files). In rare cases a signature to engine mis-match can occur that triggers a false positive (CERT, 2000). In cases like this the signature is not the cause of the false positive but simply a contributory factor. A false positive condition cannot exist in the presence of a malicious file only. Figure 1 highlights the conditions that can exist in the blacklisting approach and also shows the file types (clean and malicious) involved. This is contrasted with those conditions that can exist in the whitelisting approach where a false positive implies the incorrect classification of a malicious file as clean.

Figure 1
Conditional testing: Blacklisting and whitelisting approaches
(reproduced from Uscilowski et al., 2008)
2.2. Non-heuristic technology

A criticism of anti-virus technology is that it cannot pro-actively protect users against new unknown malicious code (Bloor, 2006). This criticism largely relates to the industry’s use of non-heuristic technology in their endpoint security products. Some characteristics of non-heuristic technology are that it is re-active, relies on vendor’s knowing and typically having access to the malicious code, and is signature or fingerprint based. Back in 1994 this technology was described as “one of the most widely-used methods for the detection of computer viruses” (Kephart and Arnold, 1994). In spite of the criticism levelled at the anti-virus industry non-heuristic technology is still a mainstay of endpoint security products today. However it subsequently has been complimented by a slew of other technologies in the meantime such as behavioural blocking and heuristic technologies (Perry, 2008). The main reasons why it is still relied upon by anti-virus vendors are because none of the technologies that have followed since have been entirely successful in making it defunct, heuristic technology is more prone to false positives (Arnold and Tesauro, 2000; Perry, 2008), and also because of the specificity it can give to malicious code. Specificity allows anti-virus software to uniquely identify a given threat or threat family (e.g. W32. Virut.CF) and is typically complimented by the provision of detailed information about the threat on the anti-virus vendor’s website (Symantec, 2009d).

Heuristics on the other hand are pro-active and are geared to detecting previously unknown viruses. They are inexact by nature and tend to be more prone to false positives (White, 1998). They can take longer to develop and even still have false negative (FN) problems (Schultz et al., 2001). Also there are many ways to develop heuristics and this label really represents a broad term for numerous technologies designed to detect new unknown malicious code.

2.3. False positives in related technologies

False positives are a well recognised problem across the information security spectrum. A number of technologies are prone to them. These include anti-virus, anti-spam and also intrusion prevention (IPS) (also known as intrusion detection or IDS) technologies. Intrusion prevention in particular has a reasonable amount of literature focused specifically on the problem (Bakar et al., 2005) (Abimbola et al., 2006) (Hooper, 2006) (Song and Kwon, 2006) (Shimamura et al., 2008). For intrusion prevention this is probably the case because “false positives have long been the bane of IDSs, leading some to even question the value of these products” (Yee, 2004). What is being implied is that IPS technology generates false positives to such a degree that it is difficult to determine what alerts are genuine true positives versus those that are false positives. This has the effect of undermining the technology in question. Some have even suggested that false positives represent a security vulnerability for IPS technology. Products using IPS technology can be deliberately and externally manipulated by controlling the generation of false positives in a process called “squealing” (Yurcik, 2002). The goal of this is to force the IPS product owner to disable the signature related to the deliberately generated false positive. This allows a real attack to enter a network uninhibited because the signature for this real attack was disabled.

With anti-spam some consider a false positive “as usually far worse than judging a spam e-mail to be legitimate” i.e. a false negative (FN) (Fawcett, 2003). In fact Ken Schneider, then CTO of Brightmail, and now of Symantec since Brightmail were acquired by Symantec in 2004, “argued that filtering even a small amount of legitimate email defeats the purpose of filtering because it forces the user to start reviewing the spam folder for missed messages” (Fawcett, 2003). This stance is interesting as the perception within the anti-virus industry is that anti-spam can afford to be more false positive prone due to the general availability of features to easily recover legitimate e-mail that has been miss-classified as spam. Data from around that period, but prior to Symantec’s acquisition of Brightmail, indicates that two leading products, SpamAssasin and Symantec’s equivalent product, report false positives of around 1% (Goth, 2003). This level of false positives is unacceptable for anti-virus technology and especially non-heuristics where the goal according to some is to have “a false positive rate as close to zero as possible” (Schultz et al., 2001).

2.4. Non-primary consideration of false positives in anti-virus research

False positives are not typically the primary focus in anti-virus related literature. One reason is that false positives have not resulted in the questioning of the value of anti-virus technology in the way that they have with intrusion prevention technologies. More often than not any discussions around false positives are in the context
of the potential cost of a new detection technique when assessing the benefits that the new technique brings in terms of combating malicious code (Briones and Gomez, 2008; Howard and Svajcer, 2008). For instance Howard et al (2008) focus on the development of new heuristic detection rules “with an emphasis on minimising false positive detections”. Likewise false positives are not the primary focus for Briones and Gomez (2008) who outline a system that automatically classifies malicious code based on its internal structure but one which “helps minimise human error and false positive detection”. Hwang and Deok-young (2007) have a similar, secondary, focus on false positives for their “Anti-Malware Expert System”.

2.5. Signature generation processes

Signatures are effectively the output of the malware analysis processes (see figure 2 below).

2.5.1. Human analysis and non-heuristic signature generation

With the growth in malicious files, today humans are responsible for analysing and generating signatures for only a small minority of the overall total. From an analysis perspective humans tend to deal with more complex and more relevant malicious code i.e. malicious code that is more likely to exist on a user’s machine. This does not always lead to complex signature development. In many cases simple signatures can be generated even for complex threats. Also signature development can be influenced not only by the malicious code being analysed but also operational requirements. For instance in order to provide a customer with protection faster, a simple signature can be created in a timely manner. Whereas a more complex signature, although more robust, would take longer to develop and would leave the customer without protection for longer. On balance though, the human analysis process is more likely than the automated processes to result in the generation of complex signatures. These more complex signatures tend to be generic and aimed towards detecting new and unknown instances of a particular threat family and are becoming more popular in terms of use by vendors today (Bureau and Harley, 2008).

The weakness of this process is that it is prone to human error that can result in a signature that causes a false positive. Only in some cases is this related to the complexity of the signature being developed. Given the concerns of false positives emanating from these human analysis and signature generation processes some have moved to develop “expert systems” to assist and aid humans. Systems like this are process based and don’t claim to be a breakthrough in terms of technology, but “a practical application of technological knowledge accumulated over many years of anti-virus research” (Hwang and Deok-young, 2007). These systems are almost a
halfway house between human analysis and fully automated processes. This is acknowledged as much by Hwang et al; “the expert system, in the end, will merely be the system that is a collection of our current automated environment” (Hwang and Deok-young, 2007). However, when not fully automated, they can act as a comprehensive aid for human analysis given the rich information they provide and essentially make the process of analysis, classification and signature generation easier and safer (Auerbach et al., 2008).

### 2.5.2. Automated analysis and non-heuristic signature generation

Today automated systems process the vast majority of incoming suspicious file streams and generate the majority of non-heuristic signatures that feed endpoint security products. This applies across the industry (Hinchliffe, 2007). The interest in this area is perhaps best epitomised by the number of anti-virus researchers who state their work focus is or was related to some form of automation. For example in the proceedings of the 2007 Virus Bulletin conference approximately 15% of all presenters/authors stated this in their biographies.

The incoming file streams to automated systems mostly comprise of collections of malicious files traded between different anti-virus vendors. Such trading, while born of necessity to keep on top of the exponential growth in malicious files, comes with some risk of false positives. Many collections traded by vendors are not comprised totally of malicious files (Gryaznov and Telafíci, 2007) (Bustamante, 2008). Often small percentages of the total are made up of clean files or files that most vendors would otherwise not detect. In relation to the clean files, whether the reason for their inclusion in these collections is due to poor processing or intentional (and perhaps bordering on a mild form of industrial sabotage) poses a risk in relation to false positives (Bustamante, 2008). The weakness of the automated approach in general is that typically it does not lend itself to addressing more than a single instance of a threat with a given signature that it generates. Conversely its strength is processing suspicious files streams in bulk and generating signatures mostly, but not always, on a one-to-one basis for each threat processed. Other criticisms of automated analysis is that it results in non-specific identification of threats and that this leads to confusion across the industry and in the user base (Gheorghescu, 2005). Also automated systems tend to not deal with threats that leverage new techniques very well as they are not geared to process them until new tools are developed and added to the automated system.

### 2.6. Clean data

The maintenance and classification of malicious code (files) is seen as a critical function of the anti-virus researcher (Bontchev, 1993). With the exponential growth in malicious code in recent years, this work that was formerly the remit of the researcher now consumes resources on a larger scale at each vendor company to the point where it puts a “strain on storage, bandwidth, process and personnel” (Gryaznov and Telafíci, 2007). Bontchev confirmed as much two years earlier when he stated “unfortunately, given the huge number of malware in existence, it is simply humanly impossible for any single person to maintain such a collection. What is needed is a reasonably large team of highly competent anti-virus researchers, employed full time in the maintenance of such a collection” (Bontchev, 2005).

What malicious code (or bad files) is to true positives, clean or benign files are to false positives (see figure 1 in section 2.1). Clean files, their collection and maintenance, are important to anti-virus vendors too. In late 2005 Marx complained, in the context of clean data, that “while well-working processes already exist in order to report new malware and add detection for those files, it is important to attain the same high quality of processes in the case of false positives. This will hopefully reduce the impact of false positives in future and we will be able to remove files causing false alarms faster than ever before”. He added that “therefore, a large collection of ‘known good’ files is essential in order to create high-quality software” (Marx, 2005). While Marx’s former comment really refers to faster incident response for false positives after the signatures causing them have been released, his latter comment indicates recognition of how clean data can be applied to improve the quality of anti-virus signatures before they are released to the user base. Other recent work discussed the need for a separate role to be created for people responsible for the creation and maintenance of clean data in an organisation (Muttik and Vignoles, 2008).

Also it is only recently that the idea that classification is as applicable to clean data as it is to malicious data has gained attention (Muttik and Vignoles, 2008). Other recent research went further and was focused specifically on this problem (Uscilowski et al., 2008). In their research Uscilowski et al. put the relative importance of clean
data into perspective in the anti-virus industry today by stating that clean data is “used in a supporting role to
test for both false positive and true negative conditions”. The importance of metadata (data that describes the
clean data) for managing and performing analysis on data sets, clean or otherwise, is also argued. Also they pro-
posed that metadata can be distinguished in two ways by its level of abstraction: file specific (lower level meta-
data) or software specific (higher level metadata). Lower level metadata can be obtained directly through automa-
ted means from the file itself (e.g. filename, MD5 hash or digital signature) and is inherent to the file itself.
Whereas higher level metadata is more difficult to derive from automated means (e.g. automatically determining
that a file was published in a certain product such as Microsoft Office 2007). While this work by Uschiowski et al.
was an “initial foray” into clean data research an area called out for further research was that of “trust” and its
importance in relation to clean data (Uschiowski et al., 2008). Another topic deemed important to clean data is its
relevance. Both trust and relevance as they relate to clean data are discussed in the following sections.

2.6.1. Trust

For some purveyors of clean data, such as SignaCert, trust is based off known provenance. They state that a
clean data set “created without a strong degree of authenticity greatly diminishes the value gained by its use”
(SignaCert, 2009). While provenance of data is critical to determining its level of trust it should not be the
only element that needs to be considered. For example software that can be proved as having been published
by Microsoft would have a high trust level according to SignaCert’s known provenance approach especially if
sourced directly from Microsoft and their release process. However this metric wouldn’t account for new knowl-
edge garnered on a file after it was published e.g. the subsequent discovery that a file released by Microsoft was
infected with a virus. This actually happened in 2002 when a file was released in the Korean version of Micro-
soft Visual Studio .NET that was subsequently determined to be infected with a virus (Microsoft, 2002). Other
major independent software vendors (ISVs) have shipped products with malicious code too including Apple and
Seagate (Marx, 2005) (Roberts, 2006) (McMillan, 2007). This calls for another element to an overall trust metric
for clean data and perhaps one based around ongoing information related to files rather than a one-time mea-
sure before files are released. So while a file could have a high trust initially based on its known provenance, its
assigned trust could change subsequently based on new knowledge learnt about that file. From an operational
perspective this would result in the clean file being revoked from a clean data set that was used for any false
positive avoidance purposes.

2.6.2. Relevance

Relevancy is another factor critical to clean data sets especially given its use in the anti-virus industry for false
positive testing. Typically the make-up of clean data sets needs to reflect what is in the user base. These have
been referred to as “targeted clean data sets”(Uschiowski et al., 2008). Relevancy includes the following factors:

• **Prevalence:** Artefacts (typically files) that are prevalent need to be included in the clean data set as an anti-
virus vendor will not want to cause a false positive that impacts millions of users worldwide. As a result their
internal clean data needs to contain such artefacts and this data set is leveraged in their quality assurance
processes (Uschiowski et al., 2008).

• **Criticality:** Some artefacts are not highly prevalent but available in commercial software that is critical to
certain users of software e.g. enterprises. For example a file from an Oracle ERP suite is not likely to be on mil-
lions of computers around the world yet if an anti-virus vendor were to release a signature that caused a false
positive condition on the file then the implications for the businesses that experienced the false positive, and
the anti-virus vendor that released the signature that caused it, could be very serious (Muttik and Vignoles,
2008).

• **Existence:** While some artefacts may have a high prevalence, a distinction needs to be made as to whether
they are still highly prevalent or not. At a given point in time a file could be highly prevalent worldwide, yet an
update by a vendor could result in that file no longer existing. If a file no longer exists in the field, even if it was
highly prevalent at a particular point in time, it is perhaps of less importance to a clean data set. This distinc-
tion has been made for malicious code already (Muttik and Vignoles, 2008) but this applies equally to clean
data even if the lifespan of a clean file is likely to be much longer than that of a malicious file.
2.7. Defect Prevention Process (DPP)

The defect prevention process (DPP) is not a software development process in itself. As Kan describes it, “it is a process to continually improve the development process”. It is a process within a process and has been utilised mostly in software development organisations. According to Kan DPP was modelled on techniques used in Japan for decades prior, based on Deming’s principles of three essential steps:

1. Collect data on failures so the root cause(s) can be determined.
2. Outline actions to prevent the root cause(s).
3. Implement these actions (Kan, 2002).

The creators of the DPP describe it as the process of “improving quality and productivity by preventing the injection of defects into a product” (Mays et al., 1990). Troster et al. describe it similarly and recognise “that it establishes a framework for continuous improvement of products and the processes”. Troster et al. also state some limitations with DPP particularly that some elements of it don’t scale very well e.g. management of information related to the process (Troster et al., 1993).

The process was first formally implemented at the IBM Communications Programming Laboratory at Research Triangle Park, North Carolina in the mid-1980s. Prior to the adoption of the process, and before the advent of software engineering, defect detection was a key focus of programming development processes. However even with the coming of software engineering and with defect prevention as the main concept of interest, defect prevention techniques were limited and ultimately this approach was being under-exploited (Jones, 1985).

2.7.1. Process embedded

The DPP is a real-time process and is embedded within the software development lifecycle at all points. It exists all along the main path of the development cycle (Mays et al., 1990). However this suggests that the software development lifecycles are of sufficient duration for a process like the DPP to be effective. In the case of IBM it would seem projects have a timeframe of months or even years. Where software development lifecycles are short, and can be measured in hours and minutes, such as the signature generation lifecycles in the anti-virus industry, this poses challenges for leveraging the DPP as is. However others, not faced with such process time constraints, choose to use elements (causal analysis) of the DPP differently and in the form of a “retrospective and ‘end of development’ feedback process” (Leszak et al., 2000). This is a form of post-mortem analysis.

2.7.2. Organisational considerations

The action team, according to the DPP, the team created to implement preventive action, should range in size from three to four members in a software development organisation of 20 to 50 people. In a larger organisation of 200 to 250 people an action team of eight to ten members is appropriate. However this applies only to an organisation or part of an organisation that shares the same software development process. “For a product development area, a systems test organisation, and an information development organisation, three separate action teams would probably be required” (Mays et al., 1990).

The emphasis of the DPP is on the direct involvement of programmers in improving the process. That being said there is a critical role for management to play. May et al. see this role as having four parts:

- Top down support for the defect prevention process
- Provision of adequate resources to provide it with every chance of success
- Authorise the action team to improve current processes with preventive actions
- Monitor the process on an ongoing basis to ensure its continued effectiveness (Mays et al., 1990).

More recently others have focused on organisational impacts of aspects of the DPP such as root cause analysis. This work centred on defect causes in products developed by virtual teams. One of its key findings was that “communication”, as a category, had the highest number of root causes (Jacobs et al., 2005).
2.7.3. Successes
Initially developed at IBM in North Carolina, the DPP was implemented in more than twenty-five organisations at seven IBM locations. It was also added to IBM’s internal training curriculum and delivered to developers at every major programming centre within IBM. From a product perspective, IBM’s Network Communications Program had a 54% reduction in error injection during development and a 60% reduction of in-field defects after DPP was adopted. Also, IBM in Houston, Texas, developed the space shuttle on board software control system with DPP and achieved zero defects since the late 1980s (Kan, 2002). Others across industry have recognised the benefits of defect prevention and have added to the body of research (Chillarege et al., 1992) (Moll et al., 2002). Chillarege et al. call out a limitation of DPP as being qualitatively driven, and pursue an approach that lends itself to measurement and quantitative analysis, while the Moll et al. combine DPP with other SPI elements like defect classification and defect detection. A contributory factor to the interest in DPP is that it can be applied to any development process (Kan, 2002).

2.8. DPP adoption in CMMI
A software process (SP) represents “the core processes, undertaken by a software house/department to plan, design, build and test their software products” (McCabe, 2006). A Software Process Improvement (SPI) initiative is undertaken by an organisation to improve their current SP principles and activities to ensure that the organisation is successful in its pursuit of quality so it stays competitive in business (McCabe, 2006). Two internationally recognised standards are ISO 9001:2000 (International Organisation for Standardisation) and CMMI. The former is generic and applies to any discipline whereas the latter focuses on software engineering and related disciplines (Pauk et al., 1993). The importance of defect prevention in the context of the CMMI is discussed in the following section.

2.8.1. CMMI and defect prevention
The CMMI defines five levels of maturity from Level 1 (Initial) to the most mature level, Level 5 (Optimising). CMMI-DEV version 1.2, specific to product and service development processes was released in August 2006, and superseded CMMI for systems engineering and software engineering (CMMI-SE/SW). According to the Software Engineering Institute (SEI) the purpose of CMMI for Development “is to help organisations improve their development and maintenance processes for both products and services” (SEI, 2006).

Level 5 (Optimising) is the highest maturity level in the CMMI. One of the two “support process areas” at maturity Level 5 is called “Causal Analysis and Resolution”. This area involves the following:
- “Identifying and analysing causes of defects and other problems”
- “Taking specific actions to remove the causes and prevent the occurrence of those types of defects and problems in the future” (SEI, 2006).

The CMMI also states the following: “reliance on detecting defects after they have been introduced is not cost effective. It is more effective to prevent defects from being introduced by integrating causal analysis and resolution activities into each phase of the project” (SEI, 2006). In short the core areas of DPP are one of two criteria (or support process areas) that differentiate the highest CMMI maturity level, Level 5, from the next highest, Level 4.

2.8.2. CMMI Level 5 successes
The CMMI has been used widely since its first release in 1993. At Motorola where CMMI was deployed quality increased significantly as a project advanced through the CMMI levels. Motorola found that the defect injection rate decreased by approximately half as a project moved up each maturity level. This means that a project operating at Level 5 would have a defect injection rate eight times less than a project working at Level 2 (Diaz and Sligo, 1997). Working at the higher levels represents huge cost savings when the cost per defect removal is applied. However Motorola noted that achieving higher levels of maturity requires investment throughout the lifecycle and especially the integration of tools into various processes to collect and then help with interpretation of quantitative data (Diaz and Sligo, 1997). More recently Boeing’s Space Transportation System Software division recorded impressive results as they moved to Level 5 over less mature levels:
2.9. Defect cost relative to identification point in process

It is widely documented that the later in the software development lifecycle a defect is found the more expensive it is to address (Boehm and Basili, 2001) (Shull et al., 2002) (Kan, 2002) (Galin, 2004). Empirical evidence is also available to support this. Boehm and Basili observed that finding and fixing a problem once software is released to the field is often 100 times more expensive than finding the defect at one of the earlier stages in the software development lifecycle such as during the requirements phase. In addition finding defects earlier in the process leaves fewer defects to be found and repaired later (Leszak et al., 2000). Also according to Galin (2004) “professionals agree that the proportional costs of defect removal have remained constant since the surveys conducted in the 1970s and 1980s”. Table 1 below, adopted from Galin (2004), shows the representative average relative defect-removal costs based on Boehm (1981) and Pressman (2000).

<table>
<thead>
<tr>
<th>Number</th>
<th>Software development stage</th>
<th>Average relative defect removal cost (cost units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Requirements specification</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Design</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>Unit tests</td>
<td>6.5</td>
</tr>
<tr>
<td>4</td>
<td>Integration tests</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>System tests/acceptance tests/system documentation review</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>Operation by customer (after release)</td>
<td>110</td>
</tr>
</tbody>
</table>

Also more recent research (Shull et al., 2002) differentiates defects by severity level and applies defect-removal cost based on this differentiation. Interestingly the cost of defect removal differs enormously between the two severity levels defined (i.e. low severity and high severity defects). Specifically their results stated the following:

- “Finding and fixing a severe software problem after delivery is often 100 times more expensive than finding and fixing it during the requirements and design phase.”
- However “finding and fixing non-severe software defects after delivery is about twice as expensive as finding these defects pre-delivery.”

Given the difference in costs between these two types of defect it makes sense for organisations to firstly classify their defects across severity lines and then to prioritise their resources accordingly so that the prevention and detection of higher severity defects is paramount from a quality assurance perspective.

2.10. Summary

In summary the key points from the literature review are:

- A false positive is the erroneous detection of a clean file.
- For a false positive condition to exist in the blacklisting paradigm a clean file must be present.
- Non-heuristic technology is characterised as being re-active and for use in response to known threats. This technology has been a mainstay of the anti-virus industry since its inception.
- False positives are a secondary focus when featured in anti-virus research.
- There are three critical stages to the signature generation process. The first is the incoming source of suspicious files, the second is the analysis and classification of those files and the final stage is the generation of...
signatures (if any of the files were classified as malicious) and their subsequent release.

- This signature generation process can be automated or manual. Both can make errors that can result in false positives.
- Clean data is a critical component in terms of false positive avoidance.
- Important considerations of clean data are trust and relevance. Trust is based off provenance but is not the only factor that needs to be considered. Relevance considers factors like file prevalence, criticality and whether the file still exists in the user base or not.
- DPP is a process geared to preventing the injection of defects into the software process. It was developed by IBM and it has been leveraged by industry standards such as CMMI, specifically CMMI’s highest maturity level (Level 5).
- A core aspect of DPP is root cause analysis of defects so that preventative measures can be taken and implemented to address the causes.
- DPP has had success at many organisations since it was proposed and implemented at IBM in the mid 1980’s. Some organisations that have benefited from it include Boeing and Motorola.
- The cost of removing and fixing a defect is more expensive (up to 100 times in certain cases) the later in the software development lifecycle it is found.
- Other research has found that the cost of removing a defect depends on its severity level. Finding and fixing higher severity level defects is 100 times more expensive when that defect is found in-field than when found during the requirements stage. For lower severity defects the cost is about twice as expensive.
3. Methodology

This chapter outlines the approaches considered for the research. The approach ultimately selected is discussed and justified.

3.1. Research question and objectives

The main goal of this research is to establish a framework that leverages software process improvement (SPI) techniques to eliminate high severity false positives from non-heuristic signatures released by Symantec. In addition this should be achieved in a cost efficient manner while also having a positive effect on the overall number of false positives from non-heuristic signatures. The framework should also provide visibility to management and engineers of signature quality throughout the content generation lifecycle.

3.2. Approaches considered

At the outset a number of research approaches were considered. The literary review showed that there was a lack of academic research focused on quality assurance problems facing the anti-virus industry. This meant there was no precedent that could provide direction on a suitable methodology for this research (Remenyi and Money, 2004).

A research approach considered was the use of large-scale surveys (questionnaires). One of the main reasons for adopting this approach is the ability to collect a large amount of information from a sizeable population (Remenyi and Money, 2004). This wasn’t appropriate for this study as a large population with the required knowledge did not exist. The next approach vetted was the case study. This approach appeared to lend itself best to answering the research question. This is covered in the next section. The philosophical approach for this research is inductive (Saunders et al., 2007).

3.3. Case study

After careful consideration the case study is the research strategy that was chosen to assist with answering the research question. According to Saunders et al. (2007) the “case study strategy is most often used in explanatory and exploratory research”. The research for this dissertation has both an explanatory and exploratory focus. The explanatory focus relates to studying a particular problem in order to “infer causal relationships between variables”. The exploratory focus relates to seeking new insights into a phenomena (such as false positives) and also the author’s role as a “practitioner-researcher” (Saunders et al., 2007). Remenyi and Money state that “the aim of the case study is to provide a multi-dimensional picture of the situation” (Remenyi and Money, 2004). Key characteristics of a case study are the focus on a particular organisation, the historical importance and context of the topic with respect to that organisation, and also the ability and need to triangulate different sources of data. Also a variety of data collection techniques are typically used (Saunders et al., 2007).

According to Yin there are also two dimensions to consider for case studies:

- Single case v. multiple case.
- Holistic case v. embedded case (Yin, 2003).

In the first dimension a single case represents a “critical, extreme or unique case”. On the other hand it may be utilised “because it is typical” or when it provides an opportunity to analyse “a phenomenon that few have considered before”. Multiple cases are used when the results from various cases need to be compared to see if the results from each are common and can be generalised.

The second dimension relates to the unit being researched. Conducting the case study on a company, or a unit in the company, where you are employed is described as an embedded case study (Yin, 2003). If the study focuses on the company as a unit then this is referred to as a holistic case (Saunders et al., 2007).

This research, according to Yin’s dimensions, can be considered a single, embedded case study. On the latter point, while the case study is focused on Symantec, only one organisation within the company produces non-heuristic signatures and so the research is generally confined to that organisation (i.e. STAR).
3.4. Practitioner-researcher role

The practitioner-researcher role is a role defined by Saunders et al. for part time students currently working in an organisation who find themselves surrounded by interesting opportunities for research. This role overcomes some key challenges of research. Some advantages of this role are:

- Knowledge of the organisation.
- Understanding of its mechanics.
- Access to data.
- Time saved in the research process due to the former points (Saunders et al., 2007).

As advantageous as this role seems, the very advantages of the role can result in structural weaknesses in the research if the researcher is not careful. For instance an outsider is more likely to ask basic questions that test important assumptions that an insider would not. An insider’s preconceptions could mean important areas could be missed that would benefit the research. Another disadvantage is that the researcher role may be difficult to differentiate from the person’s normal role, resulting in extra hours, and perhaps a lack of appreciation as to the effort required to perform the normal role and the role of the practitioner-researcher (Saunders et al., 2007).

3.5. Data types and analysis

A key part of the research is the acquisition and analysis of data. The data collection techniques for case studies can be varied and can be used in combination as will be the case for this study. Using a number of techniques both quantitative and qualitative data will be collected for this case study.

Quantitative data will be gathered from a documented secondary data source within Symantec. Secondary data is data that has already been collected for some other purpose and can be in raw or compiled form. Specifically, in this case, the source is a database belonging to the STAR organisation in Symantec. Again, such data is typically used in explanatory research (Saunders et al., 2007). Approval was sought and subsequently granted in May 2009 for use of this data for the purposes of this research. Important criteria for secondary data such as measurement validity and precise suitability are met by the identified source.

Semi-structured interviews are the source of the qualitative data to be collected. These interviews are considered non-standardised and are often referred to as “qualitative research interviews” (Saunders et al., 2007). Also this tactic lends itself to the available access of domain experts within Symantec. It also provides the flexibility to probe the interviewees to build on their answers. The analysis consists of developing a primary narrative or a summary of each interview focusing on the key themes and commonalities (Bannister, 2009). These primary narratives are in turn used to generate an overall summary or secondary, higher order, narrative and this is then is used for theoretical conjecture (Remenyi and Money, 2004).

For semi-structured interviews the researcher normally has a number of themes to be explored. Semi-structured interviews are most frequently used in explanatory research and a benefit is that they may achieve a higher response rate than using questionnaires. They are also used in explanatory research for a greater understanding of the relationship between the variables (Saunders et al., 2007).

As can be seen from table 3 above the optimal interview type for the purpose of this research according to Saunders et al. is the semi-structured interview and this was ultimately selected (Saunders et al., 2007).
3.6. Limitations

There are recognised concerns around the use of secondary data sources and of semi-structured interviews, the two data collection techniques used in this study. The main disadvantages of secondary data sources are:

- May be collected for a purpose that does not match the requirements of the research.
- Aggregations and definitions may be unsuitable.
- No real control over data quality (Saunders et al., 2007).

The first two are countered by the fact that the secondary data source in question was designed and developed to track the number and also the root causes of false positives from non-heuristic signatures released by Symantec and its data is easily accessible. However the last point is a valid concern and the risks here have been offset by control and audit measures put in place at Symantec for this secondary data source.

Qualitative research interviews have a number of data quality issues such as reliability, forms of bias, validity and generalisability. The concerns related to reliability stem from the lack of standardisation. However this in itself is a reason for some people to choose this method over others. The reason is that it provides greater flexibility especially if the interviews lead into areas of discussion that hadn’t been originally intended but which enrich the research. Bias can take the form of interviewer, interviewee and response bias. There are a number of techniques available to researchers to overcome each of these issues. Where possible these will be used. Examples include the identification of the ideal sample set and developing trust between the interviewer and the interviewee (Saunders et al., 2007).

3.7. Ethical considerations

Ethical issues concerning both methods of data collection and also the research in general were considered carefully by the author. Of particular importance were any organisational concerns regarding sensitivity of the topic and confidentiality. While the research topic could be perceived negatively as it centres on a quality problem, the focus is on developing a systematic approach to controlling the problem.

On the research tactic level, specifically for semi-structured interviews, the overarching principle is the avoidance of harm. The potential to probe and explore using this tactic means that there is greater scope for ethical issues to arise. Participation should be voluntary, preferably with informed consent, and the participants should have the right to withdraw from the process completely at any stage.

Considerations around the use of secondary data have a legal element too with the implementation of the EU directive 95/46/EC that provides protection for people related processing and storing of personal data. In relation to this study no personal data from customers is stored within the secondary data source. Beyond data collection care needs to be taken in relation to analysis and reporting so that data is not misrepresented or is selective in its presentation (Saunders et al., 2007).
4. Field Research

The goal of the field research is to collect data that will help answer the research question. Both qualitative and quantitative data will be collected. Quantitative data will assist with the root cause analysis of false positives as a result of non-heuristic anti-virus signatures released by Symantec. This will be balanced with qualitative data that will provide information on the broader context of false positives within Symantec, specific solution development, and how individual solutions can be tied into an overall framework.

4.1. Secondary data

The following sections discuss the secondary data source used for collecting quantitative data for the purposes of this research.

4.1.1. Symantec’s False Positive Logging System

The source of secondary data for this study is Symantec’s False Positive Logging System (commonly known as FPLS). The data within FPLS that is of most interest to the research is its root cause analysis data.

This system was developed in late 2006 in an effort to record data on “in field” false positives particularly those from non-heuristic signatures. The need for a system like this arose because discussions on false positives at that time were not data driven and sometimes could be emotive due to the lack of data. Also the lack of data meant that decisions in relation to solution development were made in response to a given case, rather than a population or set of cases, the result of which meant that solutions implemented didn’t necessarily address the general problem but only that specific case. A classification module was added in July 2007 so that the severity level of a false positive could be determined and recorded in FPLS. This classification module is supported by an algorithm that takes into account seven factors including impact, number of users affected, prevalence of software mistakenly detected, and the lifespan of the false positive before it was resolved. This system exists within Symantec’s Security Technology and Response (STAR) group and the Security Response Operations Quality Assurance Team are its business owners.

4.1.2. Business process

When a false positive is reported to Symantec it is typically routed to the Security Response Operations team for assessment. If the reported false positive is confirmed by this team a record is added to FPLS. Reports not confirmed as false positives are not added to FPLS. False positives get reported to this team from a variety of sources, both internal and external to Symantec, the majority through formal channels.

4.1.3. Data tracked

A variety of data is tracked in FPLS for each false positive. The pertinent data in the context of this study is root cause and severity level data.

4.1.4. Root cause

Root cause is the reason for the false positive and is selected from a pre-defined drop down list in the FPLS graphical user interface (GUI). It is the responsibility of management within the Security Response Operations team to populate this data. In most cases the root cause of a false positive is easily determined. However in a minority of cases it is more difficult to determine the reason and typically requires input from a member of the analyst/Security Response Engineer (SRE) team. Normally, the SRE who resolved the false positive is consulted first in these cases. Pertinent examples of root causes available for selection in FPLS are listed in the table below.
4.1.5. Severity level

A severity level is assigned to each false positive by the classification module. The severity level scale ranges from ‘1’ to ‘4’ and is described in Table 5 below.

Table 5

<table>
<thead>
<tr>
<th>Severity Level</th>
<th>Descriptor</th>
<th>Description</th>
</tr>
</thead>
</table>
| 1              | Minor      | • Affects a very small number of consumers or small enterprises.  
|                |            | • Occurs on file(s) with a low prevalence e.g. custom tool.  
|                |            | • Impact it normally negligible. |
| 2              | Moderate   | • Affects a small number of customers but can include some larger enterprises.  
|                |            | • Occurs on file(s) with a low prevalence.  
|                |            | • User impact is normally fair. |
| 3              | Considerable | • Affects a reasonable number of consumers and/or enterprises.  
|                |            | • Occurs on file(s) with reasonable prevalence or on file(s) that are critical to certain enterprise customers.  
|                |            | • FP can have a regional or global distribution.  
|                |            | • Often escalated to senior levels within Symantec.  
|                |            | • User impact is normally, but not always, significant. |
| 4              | Severe     | • Affects a large number of consumer and/or enterprises.  
|                |            | • Occurs on file(s) with high prevalence or criticality.  
|                |            | • FP can have a regional or global distribution.  
|                |            | • Not typically escalated as it is obvious to all that this is a critical event.  
|                |            | • Reported very shortly after signature release.  
|                |            | • Attracts negative publicity.  
|                |            | • User impact is normally severe. Can result in data loss or system instability. |

4.1.6. Data input and audit

Responsibility for inputting the various data for each record in FPLS is spread across a number of teams. These include analysts/SREs, Quality Assurance Engineers, Security Response Managers (SRMs) and functional managers (e.g. Quality Assurance Manager). The benefits of this are that a wide variety of data is tracked for each false positive and this data gets added to the system by the most appropriate team. The disadvantage is the addition of data like this is not always a priority for a given team, and when data is not added for a record, it can be difficult to track down who exactly from that team was responsible for the missing data.

Since early 2009 FPLS is audited on a monthly basis, if not more frequently, for missing records and also for missing data within records. The audit approach has been improved over this period too meaning detection of
non-compliance is high in terms of record creation. Audit capabilities for non-compliance of data input within records are developing. For instance root cause data for all false positives in FPLS was not available in approximately 2% of cases although there were no such problems when only high severity false positives were queried. However once compliance problems are identified early they tend to get resolved quickly. Importantly the two teams that can create records (SREs and SRMs) in FPLS are not the business owners of the system thus reducing any risk of measurement bias (Saunders et al., 2007).

4.1.7. Data sample set

For root cause analysis data a period from July 2008 to June 2009 was selected. This represents a sample set of 66% of the overall population in FPLS and represents the most recent data in FPLS. Also even though the total size of the sample set cannot be disclosed due to the reasons outlined in the next section the sample set is almost certainly of sufficient size so that the sampling distribution for the mean is very close to a normal distribution. For high severity false positives, due to their low overall number, the date range was not limited so that the maximum sample set possible could be obtained for analysis.

4.1.8. Limitations

The main limitation with this secondary data source is that only percentage figures can be used. If actual data on the volume of false positives was provided it could be misused, used out of context or even misunderstood. Another limitation is that this system relies on false positives to be reported to Symantec. This in itself has a dependency on a customer or third party recognising a suspected false positive in the first place and then knowing how to report that suspected false positive to Symantec. To offset this risk Symantec has developed a pro-active method for detecting these “unreported” false positives. However this study will focus only on false positives that are reported to Symantec by either internal or external parties and that exist in FPLS.

The final limitation is that the root cause has not been established for all false positive records in FPLS for the 12 month period we are interested in (July 2008 to June 2009). This only occurs in slightly over 2% of all false positive records in FPLS during this period. This limitation does not apply to any high severity false positives in FPLS. Root cause analysis has likely been attempted but these cases are more complex and require greater efforts to determine the root cause. In practice if the root cause in cases like this has not been established three to four months after the record was created it is unlikely it will ever be established.

4.1.9. Database queries

The FPLS database was queried at the beginning of August 2009. These queries provided quantitative data on all high severity false positives that existed in the database to that point and also data for a recent 12 month period (July 2008-June 2009) for all false positives regardless of severity level. The quantitative data yielded from these queries was analysed shortly thereafter. No queries were executed that involved customer data within FPLS.

4.2. Semi-structured interviews

Semi-structured interviews were selected as the method to collect qualitative data for the purposes of this research. The main reason that this tactic was selected is because the relatively small number of people who could speak authoritatively on false positives did not lend itself to other tactics. Also there was a need to collect information that neither the secondary data source nor the literature review could provide.

4.2.1. Interview preparation

The key to successful interviews is preparation and planning (Saunders et al., 2007). The following sections outline the steps taken in relation to this study.

4.2.1.1. Theme development

A number of themes were developed that would help collect data to assist with meeting the goals of the research. They were developed during late 2008 and early 2009. These themes are outlined in the table below.
4.2.1.2. Identification of participants

In terms of selection, people were approached on the basis of selecting a group of people who could provide the broadest and deepest insights into false positives from non-heuristic signatures. A mix was sought between management and technical staff and considering the context of the case study (Symantec focused) suitable experience within Symantec was preferred. Based on the criteria the people listed in the table 7 below were approached and agreed to participate in the research. There was a 100% acceptance rate.

<table>
<thead>
<tr>
<th>Interview participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Carey Nachenberg</td>
</tr>
<tr>
<td>Alfred Huger</td>
</tr>
<tr>
<td>Elias Levy</td>
</tr>
<tr>
<td>Eric Chien</td>
</tr>
<tr>
<td>Orla Cox</td>
</tr>
<tr>
<td>Kevin Hogan</td>
</tr>
<tr>
<td>Vincent Weafer</td>
</tr>
<tr>
<td>Peter Szor</td>
</tr>
</tbody>
</table>

All participants work in Symantec’s Security Technology and Response (STAR) organisation. Six of the eight are in positions more senior to the researcher. Four of the participants are members of the same team as the researcher i.e. Security Response. Six of the eight have been working with Symantec for longer than ten years with all participants having been employed by Symantec for at least five years. All interviews were requested during May 2009 with acceptance to these requests being received between late May and mid-June.

4.2.1.3. Informed consent

As part of the interview request process a range of details were disclosed to the potential participants in the form of an e-mail. The various details provided meant that the candidates were fully informed and consent to participate was freely given. This is known as “informed consent” (Saunders et al., 2007). The e-mail template used for these requests is in the Appendix but some of the key details provided are listed here:

- Purpose of the research and who is undertaking it.
- What type of data is required from the participants.
- How these data will be collected.
- How much time will be typically required to collect it.
4.2.2. Interview execution

All the interviews were conducted between 28th May 2009 and 18th June 2009. Five of the eight were held in person at Symantec’s offices in Dublin. The other three were held via telephone. While six of the eight participants typically work from Symantec’s offices in various locations in North America, three of these six were visiting the Dublin office in late May meaning three additional interviews were held in person than was originally anticipated.

4.2.2.1. Location and equipment

The location for the face-to-face interviews was a four person capacity meeting room. This provided adequate space for the researcher, the participant and the recording equipment. The interviews that took place over telephone were conducted in the same type of room as the face-to-face interviews from the researcher’s end with the participants either taking the calls from their respective office or home.

The recording equipment consisted of a laptop with audio recording software installed and a microphone which was connected to the laptop. It was tested before each interview to ensure that each interview would be recorded successfully.

4.2.2.2. Dialogue and duration

Each interview commenced with the interviewer reiterating the details provided in the interview request e-mail. In particular the right of the participant to withdraw at any stage, the preference for the interviews to be recorded, and the fact that notes would be taken during the interview by the interviewer were emphasised. The reasons provided for the last two was that it would make analysis of the data collected easier for the researcher. Once the interview started in earnest each of the themes outlined in table 6 were discussed. The discussion of each theme was typically initiated by the researcher asking open ended questions. The interviews were fluid and in all cases the interviewee was responsible for majority of the dialogue.

The average duration of the eight interviews was 46 minutes. Five of the eight lasted between 40 and 50 minutes and were close to the initial estimate of 45 minutes given to the participants at the interview request stage. The three interviews that were outside the 40-50 minute range lasted 30, 33 and 82 minutes respectively. The participant for the last one is extremely passionate about non-heuristic anti-virus technologies and their consequences such as false positives so spoke at length on each of the themes during the interview. This accounted for the interview length of 82 minutes. For the shortest two, one was simply covered in a brisk manner while for the other the interviewer was quite conscious of this person’s time so perhaps did not probe in a manner that was consistent with the other interviews. The interviewee in this case had previously made it known that they were extremely busy and could only offer up to an hour of their time for the research process, which was very happily accepted.

4.2.2.3. Scheduling

In terms of scheduling, as already stated, the interviews took place over a four week period. Four interviews were completed in the first week, two on the second week, and one each week over the remaining two weeks. While the schedule was somewhat dictated by the visit of some of the participants to the Dublin office in late May 2009, the general approach was to schedule interviews with participants that the researcher had a closer working relationship with first, and leaving the participants that the researcher worked less closely with until nearer the end of the process. This allowed time for the interviewer to develop some confidence especially with regard to the appropriate level of probing. No interviews via telephone where scheduled in the first week either.
4.2.3. Post interview tasks

Immediately following each interview the audio sessions were saved anonymously. The code for each filename that allowed for mapping of the interviewee’s name to the recording was stored separately and securely. Each recording was then replayed to assess its quality. In most cases the volume was lower than expected resulting in amplification of the recording. This editing was done with dedicated software and the resulting edited audio recordings were saved anonymously and securely. The typical size of a post-edited file for each interview was in excess of 200 MB.

During July 2009 each recording was replayed and the key points were transcribed into a summary or primary narrative (Remenyi, 2009). This took considerable time as there were over six hours of audio to listen to and often the recording would need to be stopped, rewound to a specific point so information could be transcribed, and then re-started. However the process of creating the primary narrative was made easier by using the notes that were taken during the interviews as these notes captured most of the important points from each interview. Once the primary narratives were completed the next step was to develop a higher order or secondary narrative (Remenyi and Money, 2004). This is also called the summary of summaries. This task was certainly faster than the previous one but care was taken to ensure that the key points and any commonalities across them were included in the secondary narrative. The process of reviewing the primary narratives in this manner was repeated a number of times to ensure completeness and accuracy.
5. Findings and Analysis

This chapter discusses the main findings and analysis of the case study. The first two sections (5.1 and 5.2) cover the findings and analysis of the quantitative data from FPLS while the remainder cover the findings and analysis of the qualitative data from the semi-structured interviews. The final section (5.9) summarises the findings and analysis.

5.1. Root causes for “high severity” false positives

A key goal of the research is to develop a process that aims to eliminate high severity false positives. As a result it is important to firstly understand the root causes of these false positives. “Severity 3” and “severity 4” false positives (as defined in table 5 in section 4.1.5) are normally referred to as “high” severity false positives at Symantec so this definition will be used for this research.

As stated in section 4.1.7 the actual number of high severity false positives to date in FPLS has been low so the sample set used for analysis is the entire population of high severity false positives in FPLS. Given the low numbers, the period of interest has been expanded to the maximum range of all records in FPLS and it covers a period from May 2007 to June 2009.

Figure 3 above shows the root causes of high severity false positives. The data shows that:

- 71% of all high severity false positives had a root cause of mis-analysis. Mis-analysis means a clean file entered a workflow process and was wrongly classified as a malicious file. Subsequent to that a signature was generated and released to Symantec’s customers that wrongly detected this file on their systems.
- 29% of high severity false positives had non-unique signatures as their root cause. This means that a malicious file was correctly classified. However the resulting signature was not specific enough and also resulted in clean file(s) being detected in error on user systems.

Figure 3
Root causes of high severity FPs from non-heuristic signatures (May 2007 - Jun 2009).
Source: FPLS.

As stated in section 4.1.7 the actual number of high severity false positives to date in FPLS has been low so the sample set used for analysis is the entire population of high severity false positives in FPLS. Given the low numbers, the period of interest has been expanded to the maximum range of all records in FPLS and it covers a period from May 2007 to June 2009.

Figure 3 above shows the root causes of high severity false positives. The data shows that:

- 71% of all high severity false positives had a root cause of mis-analysis. Mis-analysis means a clean file entered a workflow process and was wrongly classified as a malicious file. Subsequent to that a signature was generated and released to Symantec’s customers that wrongly detected this file on their systems.
- 29% of high severity false positives had non-unique signatures as their root cause. This means that a malicious file was correctly classified. However the resulting signature was not specific enough and also resulted in clean file(s) being detected in error on user systems.

5.2. Root causes for all false positives

Another goal of the research is to reduce the overall number of all false positives, and not only the elimination of high severity ones.
Security Response

Figure 4
Root causes of all false positives from non-heuristic signatures (July 2008 - June 2009).
Source: FPLS.

5.3. Theme 1: Historical context of false positives

This section starts the discussion of the findings and analysis from the qualitative data collected during the eight interviews. An important goal of this data is to provide contextual information on false positives from non-heuristic signatures in light of the lack of literature in this area. This is in addition to using this data to help answer the overall research question.

5.3.1. Anti-virus FPs versus FPs from related technologies

A number of the research participants stated that false positives from anti-virus technology have a far greater impact on the user base than related technologies such as intrusion prevention (IPS) or anti-spam (AS). Also the users of anti-virus technology have a more zero-tolerance policy with regards to false positives than IPS or AS. However this is related to the typical user base of each respective technology. For anti-virus technology the user base tends to be broad and includes home and enterprise users. A high rate of false positives has not been...
tolerated by these users since this technology was first used. Traditionally the IPS user base has been narrow and consisted of system administrators. These people are more technically savvy and more tolerant of IPS false positives as they have a better understanding of them. In addition because IPS has traditionally been considered a “noisy” technology, FPs were a known cost of using this technology from the outset. While AS has broad usage like anti-virus, the ability to easily resolve a false positive by the impacted user is normally a feature of AS products. Whereas the ability to resolve anti-virus false positives is more difficult due to product policy. Also the participants stated that it is hard to have a high severity false positive with either IPS or AS technology versus anti-virus technology.

5.3.2. Notable FPs

According to the participants, within Symantec, there is one false positive that stands out above all others. This is commonly referred to as the “Chinese XP FP” and occurred because of the release of a non-heuristic signature in May 2007. The root cause was the mis-analysis of clean files where two clean Operating System (OS) files from Chinese localised versions of Windows XP were classified as malicious. Technically the Chinese XP FP was two false positives, not one, as two separate signatures detected each of respective Chinese XP OS files. However as both signatures were developed and released simultaneously, and had the same root cause, this incident is almost always referred to as a case of a single false positive. In FPLS though, where the definition of false positives is applied more stringently, two false positives were recorded and classified as “Severity 4” (or severe) false positives. These are the highest classification according to Symantec classification scale for false positives as defined in table 5 in section 4.1.5.

In terms of the user impact, systems were rendered mostly unusable and the options for restoration were difficult to implement. Ironically one of the early issues encountered after initial reports started to surface was convincing people that this really was a case of a false positive. Some users mistakenly believed that the Microsoft files wrongly classified by Symantec really were malicious and that Symantec was correct to detect them. It took some time during these initial stages to convince people otherwise and that this really was a “mistake” on Symantec’s part. This added to the already difficult task of responding to this incident.

The participants stated that internally a large number of people were involved in the incident response to the Chinese XP false positive when compared with lower severity false positives. More importantly, people at all levels were involved including executive level staff whereas this rarely happens for lower severity false positives. On the rare occasion that lower severity false positives do involve executive staff it is restricted to one or two at most unlike in the case of the Chinese XP FP. In the case of the Chinese XP FP the incident response included executives travelling to China to explain the problem to the government, customers and media there. One of the participants stated that the entire incident response to this FP lasted approximately six months. These efforts, and their inherent costs, in responding to the Chinese XP false positive are supported by research by Shull et al. from section 2.9 where that work found that resolving a severe defect that was found in-field was 100 times more expensive than one found during the requirements phase (Shull et al., 2002). With a low severity false positive the incident response lasts no more than a day or two and again is supported by research that states that non-severe defects cost twice as much to fix when found in-field rather than when found during the requirements stage (Shull et al., 2002). Externally Symantec attracted a lot of negative publicity around the globe and also in China (Yang, 2007) (Keizer, 2007a, b).

The only other false positive that is of similar significance as the Chinese XP FP is one from early 2000 known as the “Blankey FP”. However this was only mentioned by a couple of participants. This false positive was not caused by a signature of any type but an unusual problem with the underlying engine that hasn’t happened since. This had a widespread impact also.

5.3.3. Impact of malicious code growth

At the outset of this document it was inferred that the problem of false positives is growing and this is related to the exponential increase in malicious code over the past few years. A number of the interview participants believe this to be largely true. Anti-virus vendors have to release larger numbers of signatures and most of the participants stated that the relationship between the growth of malicious code and false positives is linear.
Two of the participants felt another factor was at play. This relates to how malicious code looks today versus 15 years ago. Historically malware was constructed totally differently to clean or legitimate applications and if this held, according to one participant, we would rarely see false positives today. However nowadays malware is being designed to look legitimate and typically is developed in higher level languages. When combined with malware growth it exacerbates the problem of false positives.

5.3.4. Organisational attitude and customer perception

Symantec as an organisation has traditionally had an almost zero tolerance approach to false positives. One of the main reasons is that Symantec has such a large user base (120 million deployments) that any quality problems in the signatures that Symantec releases have a high chance of being exposed. A small number of the participants felt that most of Symantec’s competitors don’t have this problem. The result is that significant quality assurance efforts are undertaken by Symantec prior to the release of signatures to its customers. Other participants stated that because of Symantec’s legacy stance towards false positives it is very difficult for the company to lower the bar in terms of its standards around false positives. Some stated this was almost a competitive disadvantage as it could prevent Symantec from being more aggressive with regards to malicious code.

The participants feeling on the customer perspective was that enterprise customers hold us to a much higher standard on false positives than consumers. This is magnified if an enterprise has been affected by a false positive in the recent past (up to a maximum of two years). At the same time some commented that the population that has actually suffered a false positive is probably quite small.

5.4. Theme 2: Cost of a FP

The overwhelming response by the participants was that the cost of a false positive is very difficult to quantify. However a number did say that the cost is likely to be dependent on the severity level of the false positive. This is supported by research (Shull et al., 2002). Of note, a number stated that the cost of a false positive is less than that of other software quality issues (e.g. performance). The reasons for this are that false positives:

- Don’t persist like other quality problems.
- Tend to be short lived in general especially once identified.
- Are less common.
- Typically only impact a small portion of the user base versus other quality issues or defects.

Even if the actual cost of a false positive is difficult to quantify the participants did put forward a list of factors that could contribute to the overall cost. These are listed in the table 8 below.

| Table 8 |
|---|---|
| **Costs of a false positive** | **Description** |
| **Cost** | **Description** |
| Brand | Even customers not impacted by a high severity false positive are likely to hear about it. These FPs tend to generate a lot of negative publicity. |
| Customer Loyalty | While a false positive alone would not cause a customer to leave Symantec it could be a contributory factor. |
| Incident Response | There is a fixed cost with responding to low severity false positives and the cost is not insignificant. This happens regularly. Higher severity false positives have much higher incident response costs due to broader and lengthier involvement. |
| Reputation | The team that was responsible for the false positive may suffer damage to their reputation. High profile false positives go all the way to the CEO. |
| Sales | High severity false positives can have an impact on sales as was the case for Symantec in China in the immediate aftermath of the Chinese XP FP. |
| Avoidance | While not a “per false positive” cost, the cost of avoiding future false positives is significant in terms of resources (hardware, people etc.). One participant guesstimated that the total cost of avoidance to Symantec is $5 million annually. |
5.5. Theme 3: Legacy investment in FP avoidance

Symantec’s approach in the past to false positive avoidance was largely reactive. In other words, steps would be taken after a false positive happened to ensure that Symantec didn’t cause a false positive on the same file again. This is far from a strategic approach to false positive avoidance.

In terms of a quality assurance process, while one existed it had deficiencies. This process entailed the scanning of large corpuses of clean files. The key deficiencies called out by the participants were:

1. The vast majority of quality assurance checks were wrapped into one particular stage of the overall process or lifecycle i.e. all Symantec’s quality assurance eggs were placed in one basket. In practice these efforts comprised of a scan of a clean file corpus using the pre-release signatures. In SPI literature this essentially represents defect detection.

2. These quality assurance efforts occurred very late in the content generation lifecycle, literally just prior to that content being released to the customers.

3. The clean data utilised in these quality assurance efforts was outdated and not relevant. Even if the ideal process existed it would have been undermined by this shortcoming with the clean data.

4. Organisational responsibility for false positive avoidance wasn’t clearly defined and Symantec wasn’t organised around false positives.

The specific problems with these deficiencies are listed in the table 9 below.

<table>
<thead>
<tr>
<th>Area</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality assurance checks represented</td>
<td>• Doesn’t directly target the largest root cause of false positives.</td>
</tr>
<tr>
<td>“defect detection”</td>
<td>• Not the most efficient method of avoiding most non-heuristic false positives.</td>
</tr>
<tr>
<td></td>
<td>• Significant costs associated with hardware, storage, networking, and lab facilities to conduct scans against large clean file corpuses.</td>
</tr>
<tr>
<td></td>
<td>• Scanning clean files does not scale well.</td>
</tr>
<tr>
<td>Placement of quality assurance checks</td>
<td>• Implemented late in the overall lifecycle.</td>
</tr>
<tr>
<td></td>
<td>• Defect removal later in the lifecycle is more expensive.</td>
</tr>
<tr>
<td></td>
<td>• Customer expectations around content delivery may not be met as the signatures need to go through the entire lifecycle again if problems are detected.</td>
</tr>
<tr>
<td>Clean data</td>
<td>• Corpus in the past was small.</td>
</tr>
<tr>
<td></td>
<td>• Was not relevant with regards to clean files in the user base.</td>
</tr>
<tr>
<td></td>
<td>• Updated in a re-active manner i.e. in response to false positives.</td>
</tr>
<tr>
<td></td>
<td>• No strategic approach to clean data</td>
</tr>
<tr>
<td></td>
<td>• Low use of metadata (data to describe the clean data) resulted in poor management of clean data and operational inefficiencies.</td>
</tr>
<tr>
<td>Organisational responsibility</td>
<td>• Symantec wasn’t organised around false positives.</td>
</tr>
<tr>
<td></td>
<td>• No dedicated quality assurance team for signatures.</td>
</tr>
<tr>
<td></td>
<td>• Too much responsibility placed on the developers of the content in terms of avoiding FPs rather than on the development of an optimal process to avoid FPs.</td>
</tr>
<tr>
<td></td>
<td>• Content developers were not equipped with tools to avoid false positives.</td>
</tr>
<tr>
<td></td>
<td>• Due to the above reasons developers were more likely to defer to a FN if in any doubt about FPs ultimately impacting customer protection.</td>
</tr>
</tbody>
</table>

Overall, in the past, Symantec could say that it had a process around false positive avoidance however it was likely that it yielded a low return in terms of preventing false positives from being released.

5.6. Theme 4: Root causes

Quantitative data on the root causes of all false positives from non-heuristic signatures was put to the research participants. In particular the following two pieces of data were put to them:

1. Approximately 80% of all non-heuristic false positives recorded between July 2008 and June 2009 were caused by the mis-analysis of a clean file.
2. Approximately 10% during the same period were caused by a non-unique signature.
On the first point many of the participants were not surprised by this number. Their main reasons were that the volume of malicious code makes classification more challenging and many of the sources of malicious code that Symantec uses today are “polluted” with clean files. This is a recognised problem across the industry (Gryaznov and Telafici, 2007). In addition malicious code today is designed to look legitimate.

On the second point some thought the figure would have been slightly higher. The main reasons were that more complex signatures are more prone to error. However this was countered by one participant who pointed out that it is difficult to have a high percentage of false positives from non-unique signatures when you don’t release high volumes of the more complex type signatures that cause them. In other words lower complex signatures are less likely to have a non-unique signature as a root cause.

5.7. Theme 5: Ideal approaches to FP avoidance

The participants were asked what an ideal approach to avoiding false positives would look like given the underlying root causes. The core building blocks that they discussed for an ideal approach are encapsulated in table 10 below.

<table>
<thead>
<tr>
<th>Components of an ideal FP avoidance approach</th>
<th>Reasons</th>
</tr>
</thead>
</table>
| Application of defect prevention            | • Triggers much earlier in the lifecycle.  
                                           | • Consumes fewer resources.               
                                           | • Specifically all incoming sources of files to Symantec should have a clean file filter applied to them so they are left with a higher percentage of malicious files.  
                                           | • However defect detection techniques are still required as defect prevention is not effective against the causes of all FPs. |
| Quality assurance throughout the lifecycle  | • Quality assurance should be applied throughout the content generation lifecycle.  
                                           | • Assurance measures need to be applied at the optimal “injection points” in the overall process. |
| Clean data                                  | • Pro-active approach towards clean data collection and management.  
                                           | • Focus on relevant files as a priority (e.g. highly prevalent or critical files).  
                                           | • Should be used throughout the lifecycle.  
                                           | • Greater use of metadata with regards to clean data so it can be classified, analysed, and managed. |
| Communication                               | • Many teams are involved in the signature generation process, across many different locations. Channels for communication about FPs should be opened between them and encouraged. |
| Establishing a dedicated quality assurance team | • Assess quality from the entire signature generation lifecycle (not just a certain stage).  
                                           | • Broad view of quality (not simply testers).  
                                           | • Ensuring it reports into a “neutral” director and not a development director.  
                                           | • Responsible for tracking and analysing in-field false positive data including root cause data. |
| Tools/Training for Developers (Analysts/SREs) | • Adequate tools needs to be provided to the analysts/SREs to help avoid FPs.  
                                           | • Adequate training on how to prevent false positives needs to be provided. |
| Product policy changes                       | • Need better ways for our anti-virus based products to recover from false positives.  
                                           | • Means a lessened impact when FPs do occur. |
| Use of in-field technology                  | • Utilise technology to lessen the number of users impacted by a false positive.  
                                           | • Utilise technology to suppress certain in-field FPs e.g. digital signature checking prior to detection.  
                                           | • Utilise white-lists internal to the product to avoid scanning known clean files.  
                                           | • Utilise technology so signatures can be tested in-field without any impact to the user. |

5.8. Theme 6: Future FP trends and risks

Risk of future false positives from non-heuristic signatures is dependent largely on how relevant non-heuristic technology will be in the future. In the short term most participants stated that non-heuristic technology will be very important to Symantec and one that the company can’t do without. The main reason is that most heuristic
based technologies are still quite immature. However in the longer term most of the participants stated that non-heuristic technology is likely to be simply one anti-virus technology of many used within endpoint security products. Most participants stated that its relevancy will decline over time although two participants did state that non-heuristic technology will not be as irrelevant in the future as some people think.

Overall the participants stated that the risk of false positives is likely to shift to heuristic based technologies. One commented that some competitors have shown higher true positive (TP) rates with heuristic technologies although these have been more prone to false positives. Interestingly the participant stated that problems like this don’t appear to have impacted the competitor’s market share so perhaps end users are not as sensitive to false positives as we think. A number of participants stated that non-heuristic technologies will result in reasonable numbers of lower severity false positives in the future and while great efforts will be made to prevent high severity false positives Symantec will likely not be completely immune from them.

5.9. Summary

In this chapter the root causes of non-heuristic false positives of higher severity false positives and also all false positives regardless of severity level were discussed. In addition qualitative data from semi-structured interviews was also detailed. Figure 5 represents the signature generation lifecycle at a high level and provides a representation of the most important findings from this chapter.

Figure 5

Representation of main findings and analysis.

(Parsons 2009)

Stage 1: Pre-signature development

- The root cause of 71% of high severity FPs originate here i.e. clean files incoming to Symantec are mis-analysed.
- The root cause of 81% of all FPs, regardless of severity level, originate here. The cause here is also mis-analysis of a clean file.
- However there was little emphasis (in terms of focus and investment) in this area in terms of FP avoidance i.e. defect prevention.
- Little to no use of clean data at this stage.

Stage 2: Post signature development (pre-release)

- The root cause of 29% of high severity FPs originate in this stage i.e. a non-unique signature is created for a malicious file but this also detects clean fields in the user base after it is released.
- The root cause of 12% of all FPs, regardless of severity level, originate here.
- However the vast majority of investment in FP avoidance was applied at this stage i.e. defect detection measures.
- Also the QA measures implemented happen very late in stage 2 just before signature release.
- Resulted in greater likelihood of defects being detected later in the lifecycle and also higher removal costs.
- Clean data utilised here to detect defects was not relevant and of poor quality, meaning defects were less likely to be detected.
- No strategic approach to clean data.
Figure 5 divides the overall lifecycle into three stages and applies the key information yielded from the research to each stage. In the figure above there are many factors that do not apply to a given stage, and apply to the whole process such as:

- Being organised around false positives.
- Having a dedicated quality assurance team in place.
- Ensuring communication channels exist and are being utilised.
- Measurement tools are deployed for data extraction and interpretation at all stages of the lifecycle.

However the figure may still afford an easy way to interpret many of the key findings from this chapter.

Overall one could argue that the key finding is that 71% of all high severity false positives (and 81% of all false positives regardless of severity level) from non-heuristic signatures, are caused by the mis-analysis of a clean file that enters one of Symantec’s content workflow systems. This software failure has its origins at the very early stages of the content generation lifecycle and before the file has been analysed, wrongly classified, and the offending signature created.

Another key finding is that the majority of legacy investment to avoid false positives is performed very late in the lifecycle (late in stage 2 in figure 5), just prior to signature release. As discussed earlier (see section 2.9) the removal of defects late in the lifecycle incurs far higher costs than at an earlier stage. These costs are not insignificant to Symantec.

The final key point is to do with clean data. How clean data is sourced is critical to avoiding false positives. In the past clean data was collected re-actively and without the support of strategic planning. Also, once solicited clean data was not managed optimally because of its lack of metadata. Even if Symantec had the ideal process in place to avoid false positives, it would have been undone by these problems with its clean data.
6. Conclusions and Future Work

The overarching goal of this research was to establish a framework that leverages software process improvement (SPI) techniques to eliminate high severity false positives from non-heuristic signatures released by Symantec. In addition this should be achieved in a cost efficient manner while also having a positive effect on the overall number of false positives from non-heuristic signatures. The framework should also provide visibility to management and engineers of signature quality throughout the content generation lifecycle.

6.1. Framework development

Each of these framework requirements described above have been examined in course of this study. This was done using a combination of techniques such as a review of the existing literature, utilising a valid source of secondary data at Symantec, and also through the analysis of primary data solicited directly from domain experts at Symantec. This yielded the following information that is relevant to the development of a framework:

• 71% of high severity false positives were as a result of mis-analysis. The equivalent figure for all false positives was 81%.
  ° This was the cause of Symantec’s most notable false positive this decade, the Chinese XP FP from May 2007.
  ° A solution to this would be a clean file filter injected into the earliest stage of the signature generation lifecycle.
  ° This solution focuses on defect prevention rather than defect detection and is supported by extensive quality assurance theory (i.e. DPP and CMMI). Empirical data from a number of studies (see section 2.9) shows that the cost of removing a defect is more costly the later in the lifecycle this is done. Conversely it is cheaper to remove defects earlier in the lifecycle, again supporting this solution.

• 29% of all higher severity false positives were as a result of a non-unique signature. The equivalent figure for all false positives was 12%.
  ° Symantec’s legacy approach to false positive avoidance was geared towards defect detection.
  ° The approach is costly to operate and does not scale well.
  ° This approach likely existed due to other deficiencies such as organisational ones that included the lack of a dedicated quality assurance team to focus on the broader process. These deficiencies meant that for many years no efforts were made to determine the root causes of false positives. With little data there was no basis to question this legacy approach to avoiding them.
  ° Defect detection was, and still is, a valid approach for testing for false positives where the root cause is a non-unique signature. However the fact that this root cause accounts for 29% of high severity false positives and 12% of all false positives needs to be taken into consideration in relation to future investment in false positive avoidance.

• In the past Symantec was not organised around false positives.
  ° No single team was officially responsible for false positives.
  ° Too much responsibility was placed on the developers of content leading to a risk aversion approach with regards to tackling malicious code rather than on emphasis of quality assurance throughout the process.
  ° A solution is to create a dedicated quality assurance team that is focused on the overall process and lifecycle. In addition the organisation needs to focus on a process based approach to avoiding false positives rather than the legacy approach which focused on particular stages of the process in isolation.
  ° This team should also be responsible for the extraction, interpretation, and presentation of false positive data from tools placed at critical points in the signature generation lifecycle.

• Clean Data
  ° Clean data wasn’t relevant for its role in FP avoidance in the past.
  ° There was no strategic approach to clean data with regard to its critical role in helping avoid false positives. The legacy approach was re-active and infrequent.
  ° A lack of metadata led to poor management of clean data and to inefficiencies when it was used operationally.
  ° Solutions include creating a clean data strategy to ensure that the relevancy of clean data is paramount and that re-assessing this on an ongoing basis is critical too. Also ensuring the availability of metadata allows for greater re-use, analysis, and operational efficiency of clean data.
6.2. Framework outline

Using this data a proposed framework has been outlined in the figure 6 below. Like figure 5 it does not include factors that apply to the entire process such as:

- Being organised around false positives and support from senior management.
- A strategic approach to clean data with adequate/ongoing investment.
- A dedicated quality assurance team in place that reports into a neutral director and is mandated to consider quality across the entire lifecycle and not just at one particular stage.
- Ensuring communication channels exist and are being utilised especially as there are many teams involved in the signature generation lifecycle and that they are located across many different sites globally.
- Measurement tools are deployed for data extraction and interpretation at all stages of the lifecycle.

These factors can be considered layers to be applied to the underlying framework.

Figure 6
Framework outline for preventing false positives from non-heuristic signatures at Symantec (Parsons 2009)

<table>
<thead>
<tr>
<th>Stage 1: Pre-signature development</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Substantial investment in this stage of the process to directly target the cause of 71% of high severity false positives i.e. the mis-analysis of a clean file.</td>
</tr>
<tr>
<td>• Substantial investment in this stage of the process to directly target the leading cause 81% of all false positives regardless of severity level i.e. the mis-analysis of a clean file.</td>
</tr>
<tr>
<td>• Focus is on defect prevention.</td>
</tr>
<tr>
<td>• Solution manifests as a clean file filter (metadata based) that gets applied to all incoming suspicious file streams.</td>
</tr>
<tr>
<td>• Defect prevention is a cost efficient way of preventing defects from being injected into the overall lifecycle.</td>
</tr>
<tr>
<td>• Filter needs to be driven by trusted sources of clean data called out in the clean data strategy. Also the clean data used needs to be relevant.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 2: Post signature development (pre-release)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Continued but reduced investment at this stage relative to the overall investment in avoiding false positives.</td>
</tr>
<tr>
<td>• Investment should be influenced by the percentage of root causes that originate during this stage i.e. 29% of high severity false positives and a maximum of 19% of all false positives regardless of severity level.</td>
</tr>
<tr>
<td>• However the declining influence of non-heuristic technology may mean the significance on this stage of the process may increase over time.</td>
</tr>
<tr>
<td>• Focus is on defect detection.</td>
</tr>
<tr>
<td>• As defect detection measures for false positives does not scale well better management of the clean data used at this stage is required so that resources are use efficiently.</td>
</tr>
</tbody>
</table>
6.3. Framework implementation at Symantec

Some elements of this framework have been in place for some time at Symantec as they were part of the legacy approach to avoiding false positives. Other non-legacy elements have been implemented more recently. Arguably the most important element within the framework has only been implemented since the start of this year (2009). This is the clean file filter and it is applied to all files incoming to Symantec’s analysis and classification systems. Since the start of the year, on a monthly basis, it filters approximately 2-2.5% of all incoming files. In March 2009 approximately 60% of the clean files filtered were Microsoft files. An instance of mis-analysis on a Microsoft file has a high possibility of being classified as a high severity false positive demonstrating the benefits of the clean file filter. In essence each of these clean files filtered represents a “potential” false positive as if they were not filtered they provide an opportunity to a human or automation for mis-analysis. If each of these files was not filtered it would not necessarily result in a definite false positive, it merely increases the chances of it. It is for this reason that the actual number of false positives that did not manifest due to the application of the clean file filter is difficult to quantify. Although recent data from FPLS shows that Symantec hasn’t recorded a high severity false positive (i.e. severity 3 or 4) in over 13 months (figure 7 below). However as the clean file filter was implemented halfway through this period it can’t be entirely responsible for these improvements. Other elements of the framework that have been implemented have played a role also e.g. tools to assist the analysts/SREs to avoid false positives and the application of more relevant clean data throughout the signature generation process. That being said the volume of signatures released by Symantec has increased by 55% in the six month period from January 2009 to June 2009, over the previous six month period from July 2008 to December 2008, suggests that the clean file filter is providing benefits in terms of preventing higher severity false positives.
An incidental benefit of the clean file filter is that by filtering up to 2.5% of all incoming suspicious files, means that such files are not subject to classification processing (automatic or manual). This equates to thousands of files not having to be assessed by hand every month.

Finally what can be inferred from this research is that some false positives from non-heuristic signatures may actually be acceptable, as the cost of addressing them is low, while the cost of eliminating them completely is likely to be excessive. However this only applies to lower severity false positives. Higher level false positives are deemed unacceptable to both customers and Symantec and should be avoided at all costs.

### 6.4. Generalisability

While the focus of this research is specific to Symantec, the one technology that pervades the industry more than any other is non-heuristic technology. Non-heuristic technology is the one that differs the least between vendors in addition to being the most commonly used. The result is that while this research will not apply 100% to other vendors, it should have broad applicability.

### 6.5. Advancement of current thinking on topic

The goal of most research is to make a significant or valuable addition to the body of knowledge (Remenyi and Money, 2004). This research aims to provide dedicated research on the causes of and solutions for false positives within the anti-virus industry. The research is focused on a mature technology (non-heuristic) that is used widely across the industry.

The fact that Symantec has used techniques such as root cause analysis and leveraged broader SPI theory like DPP may be of interest to people within the industry. Also how Symantec has applied these techniques to an environment that perhaps they were not envisaged for may be of interest to quality assurance practitioners outside...
of the anti-virus industry e.g. an environment with very short software development lifecycles that are typically measured in minutes and hours rather than months or even years as was the case at IBM (Mays et al., 1990). It shows that even if quality assurance standards or theory aren’t adopted wholesale by the anti-virus industry, leveraging important elements of them may yield benefits to both the industry and its customers.

6.6. Limitations

The interviews were limited by not holding follow-up sessions with each participant. This would have allowed the participants time to reflect on themes and the questions asked and subsequently to provide a more complete account and ultimately more accurate data. Nevertheless the data solicited from the set of interviews that were held did provide data that enriched the research process.

The secondary data is limited by the fact that not all false positive conditions that manifest in the field get reported to Symantec. It cannot be determined what impact this could potentially have on the quantitative data used in this study.

6.7. Future directions of research

While non-heuristic technology is mature, approaches for dealing with false positives emanating from the technology are not fully developed. Work building upon this study is an obvious starting point for future research. The lack of literature in this area does beg the question as to why there is a lag between the development and the subsequent maturity of a technology, such as non-heuristic, and dedicated research on quality assurances issues related to it. As one participant in the interview process pointed out, the focus at industry conferences is around anti-virus researchers with minimal participation by quality assurance researchers thus the research rarely has a primary focus on quality assurance issues. There may be many reasons for this though. However these factors need to be considered when focusing on the problems of current research in this area such as:

• In the literature review quality assurance is often a consideration of the research but it is rarely the focus of it.
• Quality assurance is often discussed as being synonymous with testing perhaps demonstrating a lack of understanding of quality assurance theory within the industry (Muttik and Vignoles, 2008).

The following areas related to false positives are potentially interesting avenues for future research:

• Pan-industry research on false positives from non-heuristic signatures. This could also test some of the findings specific to Symantec from this research i.e. its generalisability.
• Research on false positives from all anti-virus technologies at Symantec. One of the very reasons, although there are others, why the likes of heuristic and behaviour blocking have not supplanted non-heuristic technology is due to their higher false positive rates so research in this area would be of interest too. The same applies to reputation based technologies but these are just in their infancy and may need to demonstrate their value before it would be worthwhile performing research on false positives in this area.
• Clean data is the critical ingredient to any false positive prevention infrastructure on an ongoing basis after its development. Clean data was discussed to a limited degree in this research relative to the scope of discussion that it could entail. Further research could be pursued related to its classification, relevancy, and cost of acquisition amongst others.
References


Appendix

E-mail request to potential research participants

Hi xxxx,

I’m currently working on a research problem for my dissertation and I’m really hoping you can help out.

The dissertation is a case study on False Positives from non-heuristic signatures and Symantec’s approach tackling this problem. As part of this research I intend to use two main data collection techniques, one quantitative and one qualitative. The qualitative one will focus around semi-structured interviews. This is where a number of themes can be discussed in an objective manner. Some realistic examples are:

• Historical FPs/FPs of note
• Costs of a FP
• Legacy investment in FP avoidance
• Root causes
• Approaches to FP avoidance
• Future FP trends and risks

As you are experienced and knowledgeable in this area I was hoping you would kindly agree to participate in an interview for the purposes of this research. Can you please let me know if you would like to participate?

The interview will typically take around 45 minutes. Ideally it will be in person and audio recorded (with your approval) from which a summary will be made (primary narrative). All summaries then are summarised themselves into an overall summary (secondary narrative) from which theoretical conjecture can be attempted. Analysis of the summary could then be used in the dissertation proper. At any stage you can decide to withdraw your participation from the process and all data solicited at that point will be discarded and will not be used. Furthermore the recordings will be anonymised when storing and no data will be attributed to the participant without prior approval. Please note the dissertation could be used at some future point as the basis for a paper to be published at a conference or in a journal.

The title of the dissertation is “A false positive prevention framework for non-heuristic anti-virus signatures. Symantec, a case study”.

If you require any further details or have any questions please let me know.

Thomas Parsons
Manager, Quality Assurance
Security Response
Symantec Corporation
www.symantec.com
Symantec is a global leader in providing security, storage and systems management solutions to help businesses and consumers secure and manage their information. Headquartered in Cupertino, Calif., Symantec has operations in more than 40 countries. More information is available at www.symantec.com.

About the author
Thomas Parsons is a Senior Manager for Quality Assurance Engineering at Symantec Security Response. He leads a global team of engineers that is responsible for the quality of content used by over 120 million systems running Symantec security software worldwide. He has just earned an M. Sc. in Management of Information Systems at the University of Dublin (Trinity College).

Copyright © 2009 Symantec Corporation. All rights reserved. Symantec and the Symantec logo are trademarks or registered trademarks of Symantec Corporation or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners.