Sit, Fetch, Drop: Training the Clearswift anti-spam filter

Emma Sutcliffe

Sit, Fetch, Drop: Training the Clearswift anti-spam filter

A Case Study

Emma Sutcliffe

GSEC Practical

Version 1.4b, Option 2.

Submitted: 26th June 2004

Table of Contents:

Table of Figures:

"Corporate e-mail users are drowning in spam. Harried IT system engineers are looking for spam filtering applications that work with other security products and that catch the most spam at the Internet gateway with very few false positives, and require minimal administrative effort."

Maurene Caplan Grey, Research Director at Gartner

Abstract

I wasn't quite drowning but was certainly tiring from treading water. Managing spam had become a daily task and I wanted a dynamic filter that could be customised to suit my environment. This case study follows my experience implementing Clearswift's Anti-spam filter for Mailsweeper for SMTP, and the challenges involved in teaching the application to adapt to my environment. As part of the fine-tuning process I evaluated the threat of spam weighed against the risk of false positives, and concluded that I would be willing to accept some spam rather than the loss of genuine email. I recognised that there is no set-and-forget solution and that successful management of spam is an on-going and dynamic process.

1.0 What we had (The "Before")

1.1 Environment

The environment is a small/medium business of approximately 450 users located across six sites. The sites are linked via various wan technologies to allow sharing of resources. There are two Microsoft Exchange 2000 Servers located behind a single SMTP gateway, which is located in a DMZ. For this case study, I have named the company Mydomain.com.

Mailsweeper for SMTP 4.3 is installed as a standalone deployment on the SMTP Gateway, GW-01, which is a dedicated gateway server.

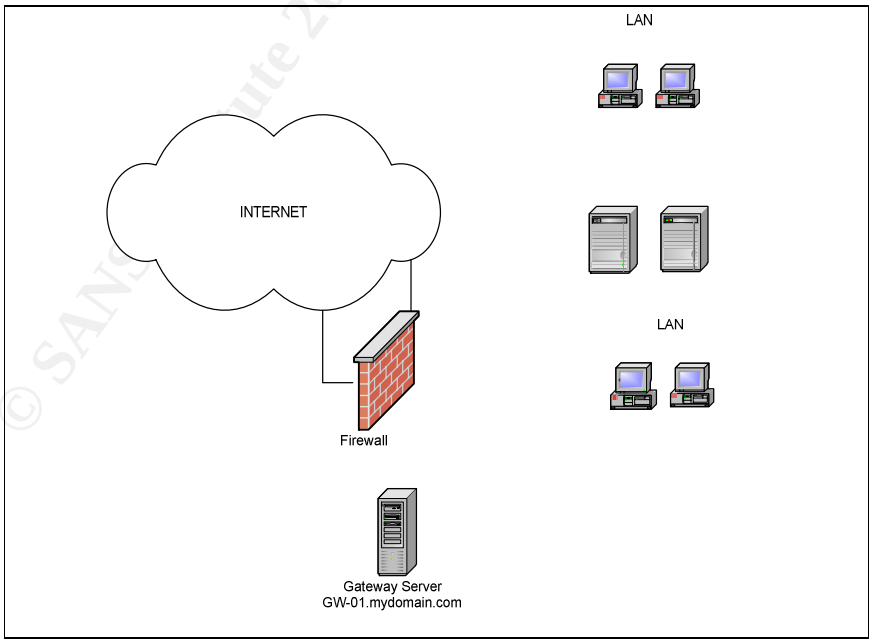

Figure 1: Simplified Network Map of Mydomain.com

On average, 4,000 – 6,000 incoming emails are processed by GW-01 per day. The existing text-based spam filter was proving inadequate in controlling an ever-increasing

volume of spam. There were numerous calls being made to the HelpDesk about the spam problem, which I estimated to include between 30-50% of all incoming email.

The following issues were consequences of the spam in our environment:

- Unnecessary bandwidth usage and congestion in mail delivery
- User inconvenience and loss of productivity
- User frustration with ineffective measures to reduce spam
- Offensive material reaching users' mailboxes
- User familiarity with spam making them less wary of more dangerous emails

1.2 Mailsweeper Configuration

While a detailed analysis of the Mailsweeper configuration is outside the scope of this discussion, a brief outline will be provided to facilitate an understanding of how the anti-spam filter will integrate with other functions performed by the SMTP gateway.

The Mailsweeper application consists of three services that work together to process the email traffic; the Receiver Service, Security Service and Delivery Service. SMTP traffic is processed by each service in turn before being passed to the next service, until the message is delivered or otherwise acted upon.

The Receiver Service validates the connecting host against the routing and relay policy.[1] If the email is from a permitted source, it is accepted and moved to a folder for collection by the Security Service. The Security Service applies the policy elements found in the Scenario Folders and assigns a Classification to the email. The Classification dictates what actions will be applied to the email and whether it will be passed on to the Delivery Service or otherwise stored in quarantine or dropped altogether. The Delivery Service then collects any approved emails and begins the delivery process to the recipient host.

Mailsweeper is installed as a standalone deployment, so all SMTP traffic is routed through the Mailsweeper server before being forwarded either internally to the Exchange servers or externally to the recipient domain.

The Mailsweeper process is illustrated in the following diagram:

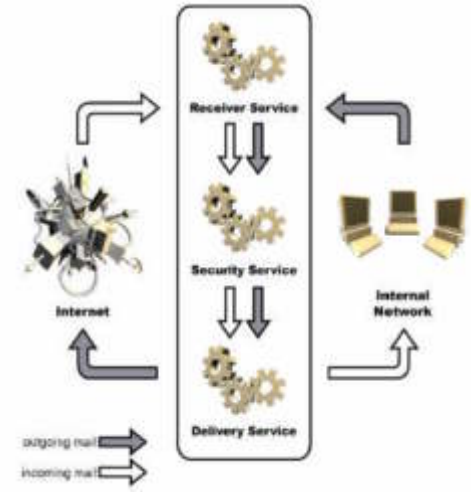

Figure 2: Mailsweeper Services
(Source: CS Mailsweeper for SMTP 4.3 Manual, page 1-10)

[1] Clearswift Anti-spam Filter Installation & Configuration Guide, page 4-3.

1.2.1 Policy creation

Different policy elements are configured as Scenarios within Scenario Folders. Each Scenario Folder contains the security policy elements to be applied to a specific set of SMTP traffic, based on its routing information.

Separate Scenario Folders have been created for Incoming and Outgoing mail, as follows:

Incoming =	from: *@*	to: *@mydomain.com
Outgoing =	from: *@mydomain.com	to: *@*

The Incoming Scenario Folder contains a variety of scenario types, including:

- Anti-Virus plug in for virus detection.
- Data Type Manager - detects types of files by file signature (not file extension), eg; executable files.
- File Detector - uses file masks to detect specific filenames, file extensions.
- Size Manager – uses a size threshold to park emails with large attachments for out-of-hours delivery.
- Reclassifier – handles encrypted, undetermined and non-RFC compliant emails.
- Classifer – used to apply classification to all emails, without further analysis.
- Text analyser – searches email text for a list of words and phrases as specified in an Expression List, eg; profanity, spam.
- And others.

The Anti-spam filter will be just one of a number of scenarios that may be triggered by an incoming email. An Anti-Spam scenario will not been created for outgoing email.

Each scenario has an associated Classification that is applied to email that meet the criteria of that Scenario. Classifications consist of one or more actions, including deliver, quarantine, log, notify and reply. Classifications are enforced on a hierarchical basis; the higher the classification, the greater priority it has.[2]

For example: An email arrives that contains two attachments: a password-protected Excel document and a program file (.exe). The email is processed by the scenario folder that was created to allow encrypted documents from a specified external domain to a group of recipients at Mydomain.com. There are two scenarios triggered by this email; Detect Encrypted Files and Detect EXE Files. In this instance, the scenario Detect Encrypted Files applies the classification *Allow Incoming Encrypted* to the email, and the Detect EXE Files scenario applies the classification *Incoming EXE*. As there is no virus threat detected, the default classification of *Clean* is also applied.

The classification *Allow Incoming Encrypted* dictates two actions: Deliver and Log.

The classification *Incoming EXE* also has two actions: Quarantine and Notify.

The classification *Clean* has one action: Deliver.

Whichever of these classifications appears higher in the list will be the one actioned. In this case, Mydomain.com's security policy states that incoming .exe files are to be quarantined and released only on the approval of an Administrator. If either of the other two classifications were to be applied, the email would be delivered in contradiction of this policy.

[2] Clearswift Anti-spam Filter Installation & Configuration Guide, page 9-26.

It is therefore important that the classification *Incoming EXE* be located higher than that of *Allow Incoming Encrypted*. The default classification *Clean* is always the lowest in the hierarchy.

1.2.2 Text-based spam Filter

The existing spam filter is based on a manually updated list of common phrases and words found in spam emails. This method has become ineffective due to the high administration overhead and the difficulty in 'keeping up' with constantly changing spam phrases. It is impractical to attempt to predict all the possible permutations of common spam phrases and words, eg: V1agr@, V|@gra, and V | AG.RA are just a few of the ways that spammers might include this word into an email.

A better spam solution is required to reduce the volume of spam being delivered, without increasing administration overhead or interfering with other policy elements.

1.3 Classifying spam

Before implementing a new Anti-spam policy, it was important for me to understand what I was going to define as spam.

While a popular definition of spam is "unsolicited commercial email", I prefer the definition of "unsolicited bulk email", as I feel it more adequately highlights a primary characteristic of spam; that being "the automated broadcast of high volumes".[3]

The definition of spam may also differ with personal perspective. What is spam to one person may be a commercial advertisement of interest to another.

At the corporate level, the anti-spam policy must consider these personal perspectives without allowing them to dictate the policy. The policy should therefore clearly define what will be considered as spam for the purposes of implementing anti-spam measures.

Email viruses and worms pretend to be from legitimate sources with believable subject lines and body text. Typically, these emails tempt the recipient to open an attachment that contains a virus. These types of unsolicited bulk emails have not been included in this analysis, as their 'raison d'être' conflicts with that of true spam.

Spam may be classified in different ways, and a different action is often desirable for different classifications. Based on the email traffic I had been monitoring, I determined 5 types of spam that would be considered for implementation of the anti-spam policy:

Type 1: Emails from hackers trying to control or steal information from users, eg:

- Emails masquerading as technical advice instructing a security patch be installed, fooling the recipient into downloading dangerous software (eg: Xombe Trojan)[4].
- Email encouraging recipients to click on a URL to confirm their bank account details, masking connection to a malicious site where malware is then accessed (eg: Bogus Banking Email)[5].

This type of spam is to be universally considered as dangerous and is to be blocked.

[3] Levitt, Mark & Burke, Brian E., page 2.
[4] Lyman, Jay.
[5] AUSCERT Alert AL-2004.10

Type 2: Scams and cons, eg:

- The son of a deposed leader from a small African Republic requesting money be transferred in order to gain safety in return for wealth.
- An institution requesting confirmation of passwords for an online account.

This type of spam is also potentially dangerous to recipients and is to be blocked.

Type 3: Unsolicited bulk advertisements trying to entice a purchase or a visit to a web site, eg:

- Advertisements for cheap pharmaceutical products and prescription-only medication where a prescription is not required
- Advertisements for pornographic websites
- Advertisements for cheap mobile phones, online degrees, stock market tips etc.

As well as being unsolicited spam, some of these types of emails may be considered offensive or be offering illegal products. However, at an individual level there may be those who are happy to receive these advertisements.

At a corporate level there is no business case for allowing these emails through, regardless of whether the recipient wishes to receive them or not. There is also an additional danger in allowing potentially offensive or illegal material into the organization; employees may consider their employer negligent in fulfilling their duty of care by failing to provide a safe working environment.[6]

This type of spam is also to be blocked at the gateway.

Type 4: Unsolicited bulk emails from a company or service that the recipient has some history with, for example:

- A user who attended a conference or subscribed to a site to download a research paper, and was automatically added to the supplier's mailing lists for promotional material, future events, market research etc.
- A user who subscribed to web sites in order to receive job vacancy notifications, view movie previews etc, now receives multiple mailings from a number of advertising partners.

Depending on the business, there may or may not be a business value in the receipt of these types of emails.

Where the spam originated from a legitimate business source, the supplier of the spam should have clearly advertised processes for the user to unsubscribe from their mailing lists. This type of email may have legitimate business use for some recipients in the organization while being considered as unwanted junk mail by others. Even if considered unwanted, the onus in this case is to be placed on the user to remove themselves from the suppliers lists, and then on the supplier to honour the request.

Blocking these emails would prevent legitimate commercial correspondence from reaching those who do wish to attend conferences or download research papers.

[6] Clearswift Spam Perspectives White Paper, page 5.

Where the email originated from a non-business activity, the organisational policy may deem that because this email is not business related, all such email is undesirable and will be blocked to all users. Alternatively, it may be allowed through and the onus placed on the user to remove their work email address from the mailing list.

These emails will not be blocked, and the responsibility will be placed on individual users to manage their email and internet browsing habits.

Type 5: Solicited bulk emails, for example

- User subscribes to receive 'joke of the day', sports updates, etc.
- User receives and forwards on chain letters and petitions that they believe to be worthy to their friends and colleagues.

While these emails are in fact not spam and may be desirable to the user, corporate policy deems them as an unwanted distraction and a mis-utilisation of resources.

Often there is a conflict in perception of how corporate policy should be applied to these types of emails. While corporate policy may be to block them, some recipients may believe it's their right and their choice to receive this correspondence.

The policy will not demand these emails be blocked, but users will be educated on what is acceptable email correspondence.

In summary, the policy to be applied is as follows:

Type 1 spam:	Block all
Type 2 spam:	Block all
Type 3 spam:	Block all
Type 4 spam:	Allow, recipients to manage
Type 5 spam:	Allow, recipients to manage

The primary risk in the implementation of the anti-spam filter will be that legitimate, business emails are trapped as spam. The aim will be to block a high percentage of spam while ensuring the false-positive rate is so low that it won't impact business.

Secondary considerations include additional load on the gateway server, and potential performance issues in mail delivery.

User education will also be a challenge. Information about how the filter works and why it has been implemented will be distributed via the company intranet. FAQ's and feedback will be encouraged while the initial training and tuning of the anti-spam policy is in progress.

2.0 What we did (The "During")

2.1 Other policy elements

Mailsweeper has policy elements that are not integrated into the anti-spam filter, but can still be employed in the prevention of spam. These will be covered briefly to illustrate the overall anti-spam policy being implemented.

(i) Real-time Block List (RBL):

In Mailsweeper Policy properties, there exists the option to check all connecting SMTP hosts against a database of known unsolicited mail senders. Mailsweeper connects to the third party website to look up the IP address before accepting the connection. In this

way, emails are dropped before they are analysed. This function is implemented within the Mailsweeper Receiver service, which reduces processing overhead as the traffic is dropped before reaching the Security service.

Because this traffic is dropped before being processed by the security service, false positives may not be detected. However, the sender does receive a notification from the blocklist organisation, to help resolve any incorrect listings.

Examples of RBLs include spamhaus.org, dsbl.org, ordb.org and a number of others.

(ii) Banned hosts and banned address lists:

Within the general policy properties are options for configuring both Banned Hosts (by FQDN or IP address) and Banned Addresses (by individual or domain addresses) lists. Any email originating from a host or address on these lists is not accepted.

These lists may provide protection against long-time spam hosts or addresses that are known to the Administrator. However, they offer limited protection against spammers who constantly change or spoof their sender address[7], or use compromised hosts from which to launch their spam[8].

The lists are manually maintained and it's up to the Administrator to update them as required.

(iii) Reverse-DNS lookup

This option is also found in the general policy properties. When selected, Mailsweeper checks connecting host's entries in DNS before accepting a connection. Mailsweeper can be configured not to accept connections from hosts that are not validated with a reverse-DNS lookup.

This may block some spam or other illegitimate email traffic, as legitimate email hosts should all check out in DNS. However, as mentioned above, spammers often use legitimate mail hosts that will pass a reverse-DNS lookup test.

(iv) Other considerations

Some spammers have been learning tricks from the writers of email worms and are spoofing the sender address to make the spam appear from a known source. I have noticed spam with a sender address of mydomain.com arriving at the gateway.

There is a simple means of blocking spoofed email by creating a scenario that blocks all mail with sender and recipient address @mydomain.com. In my environment, there is no cause for legitimate email sent from mydomain.com to mydomain.com to pass through the gateway. This may not be case in other environments, where there are possibly valid reasons for SMTP traffic from a particular domain to originate from outside that environment.

The Mailsweeper Relay and Routing properties are also configured to prevent GW-01 from being used as an open relay server. The Mydomain.com Exchange Servers are the only permitted relay hosts. While this has no impact on our spam policy, it does prevent mydomain.com from contributing to the problem.

[7] Levitt, Mark & Burke, Brian E., page 3.
[8] Leyden, John.

2.2 The Anti-Spam filter

Clearswift's anti-spam filter combines three technologies to asses whether an email is spam. These technologies are explained by Clearswift as follows:

- Bayesian Classifiers that learn about the words used in both spam and 'not spam', and then assign probabilities to each word based upon the number of occurrences.
- Heuristics that check the structure of the message to determine its 'spamminess'. As an example, whether the message is pure HTML-based as most normal email will have a text and an HTML body. Typically, only spam and newsletters will use HTML only.
- Textual Analysis which is the analysis of frequently used words and phrases as found in email subject lines and email text.[9]

2.2.1 Textual Analysis

The textual analysis is based on a list of words and phrases that can be amended and appended to by the administrator. This list of words and phrases can be searched for in the email body, subject line and SMTP header.

A weighting (as either a value between 1 -10, or simply 'detected) is given to each word or phrase that the filter detects in an email. Once the accumulated value of the combined weightings reaches a pre-configured threshold, the email is classified as Spam.

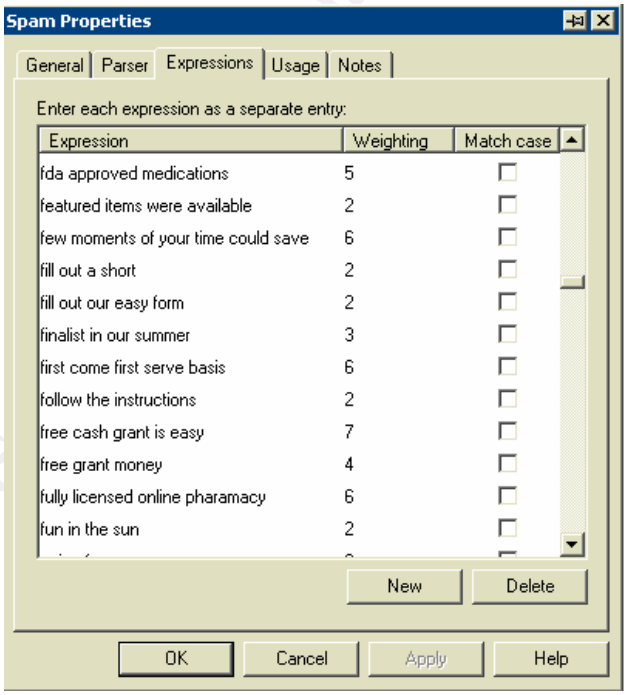

Figure 3: Sample Spam Expression List

As shown in this example, an email containing the phrase "first come first serve basis" would score as 6. If the Spam Scenario threshold is set to 10, this phrase alone would not trigger the Spam Classification. However, if the phrase "free cash grant is easy" was

[9] Clearswift Press Release

also detected in the same email, the total score would be 13 which would trigger the Spam Classification.

Clearly, spammers can bypass this type of filter through slight modifications of their phrasing. For example, the phrase “free ca.sh grant |s EA5Y” would not be detected by this particular analyser. This type of filter is frequently circumvented by newer spam, however it still provides a reliable addition to a dynamic anti-spam filter.

Certain words or expressions that the Administrator wishes to block at all times, regardless of the surrounding content, can be included in this analyser. This may include offensive words and phrases and common pornography phrases.

2.2.2 Bayesian Analysis and Heuristics

Two files are used as references by the anti-spam filter for Bayesian analysis. The first is a text file, the ASF file, which contains phrases, words and domain information together with corresponding values for each. These values express the likelihood that the word or phrase will appear in a spam email.

The second text file also contains words and phrases with corresponding values. This file is used to assign a ‘not-spam’ value to words and phrases and is known as the NSASF file. This not-spam text file is used to prevent legitimate emails from being blocked by creating a profile of words and phrases used in non-spam correspondence[10]. The non-spam value of these words and phrases is offset against the value these same words may have in the spam text file.

Example extract from a hypothetical ASF file:

> 9 mailoffers 51 4 http 10 5 100 13 3 pills 13 4 2 2 20 5 happy 20 6 24 10 3 hour 10 17 benefits 10 18 more 10 19 information 10 20 inquiries 10 3 improvement 19 4 muscular 104 4 strength 104 4 wrin-k1e 104 4 reduction 104 3 increase 19 3 emotional 104 3 additional 12 4 beauty 10 3 guide 10

Example extract from a hypothetical NSASF file:

> 2 extension 2 4 bring 2 2 it 2 3 back 2 2 all 1 3 these 1 4 things 1 5 correct 1 2 while 1 4 relating 1 5 cost 1 2 about 1 2 this 1 4 would 1 5 have 1 6 got 2 2 any 2 2 3 causing 2 4 me 2 6 happy 1 4 have 2 5 made 1 3 some 1 5 about 1 6 getting 1 2 additional 4 2 copies 2

The information in these two text files is combined and emails are analysed against the combined reference file. A value assigned to each email based on this examination of words and phrases. This value is then integrated into the results of other heuristic analysis, to determine a final score for each email. This final score translates to the likelihood that the email is spam, and thus determines how that email will be classified.

Some of the heuristic analysis that the Anti-spam filter performs on an email include:

- Number and analysis of URLs
- HTML or text email format
- Sender domain
- Subject of email
- Amount of ‘white space’ in body of email
- And others.

The combined results of these analysis techniques determine a final score. An example of the analysis results for a spam email is shown below.

[10] Clearswift Anti-spam Filter Installation & Configuration Guide, page 4-2.

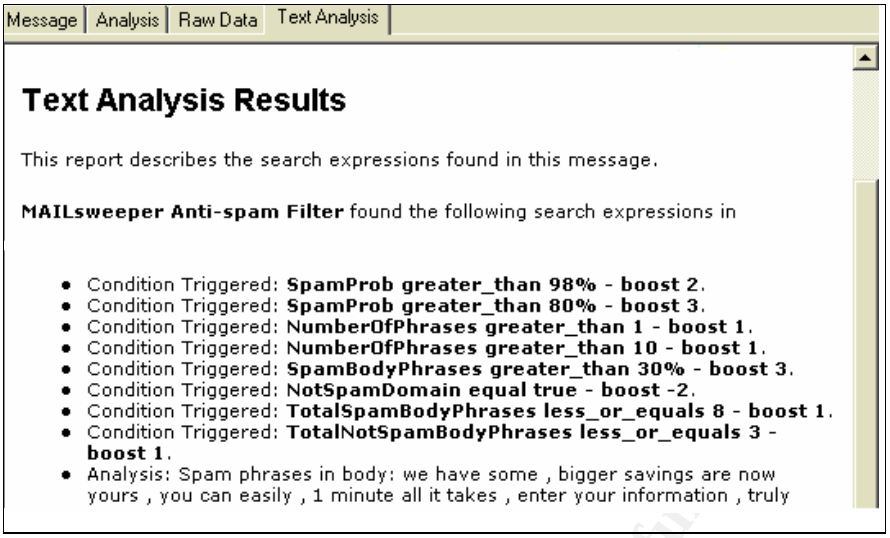

Figure 4: Example analysis results from a spam email (email ID has been removed).

Each of the criteria is evaluated and allocated a value. The value may be positive or negative depending on the results of the analysis. When all values are combined, the total score is calculated.

By default, a score of 5 or higher classifies the email as spam, and a score of 1 or higher classifies the email as 'maybe spam'.

These thresholds can be adjusted in a configuration file. For example; if too many false positives are occurring, the thresholds can be changed to 6 and 2 respectively. This means that emails scoring up to 1 will no longer be tagged as maybe spam, and emails would have to score 6 before being classified as definitely spam.

Default configuration file (thresholds in bold):

```
<classifications default="NOTSPAM">
        <classification>
                <threshold>5</threshold>
                <class>HIGH_SPAM</class>
                <modification></modification>
                <location>1</location>
        </classification>
        <classification>
                <threshold>1</threshold>
                <class>LOW_SPAM</class>
                <modification>[Maybe Spam]</modification>
                <location>1</location>
        </classification>
</classifications>
```

The configuration file allows modifications to the weightings that are applied to all criteria. If one particular criterion is trapping legitimate email, its weighting can be reduced to suit the environment.

For this installation, the default settings have been retained.

As a counter-balance to reduce the number of false positives, a 'White List' is included in the filter. The white list contains phrases, domains and URLs that the Administrator deems are not to be recognised as spam components.

That does not mean that emails originating from domains in the white list are not analysed. It does mean that the value the filter would ordinarily assign to a domain will not be included, and that domain will not contribute to the overall score.

For example: Hotmail.com and Bigpond.com.au are origin domains of much spam (as are other ISPs and free web mail providers), but these domains also serve many legitimate accounts. By including these domains in the white list, an email originating from one of these domains will be evaluated only on its other attributes (through Bayesian and heuristic analysis), to determine what its classification will be.

Figure 5: Example white list

2.3 Implementing the anti-spam filter

Installation of the anti-spam filter was a simple process of running a setup file that both installed the program and created the scenario plug-ins required to configure it. A number of utilities were included which I discovered to be useful once the training process began.

2.3.1 Scenarios and Classifications

The first big question was how and where would the anti-spam filter integrate with the other scenarios and classifications that are simultaneously being applied to incoming emails?

The two scenarios that process spam apply two classifications, one for definite spam and one for tagged or maybe spam. The classification applied to spam is *Incoming SPAM*, and the classification applied to maybe spam is *Incoming Spam TAGGED*.

I set the classification *Incoming Spam* to apply a quarantine action, and *Incoming Spam TAGGED* to apply a deliver action. I also included a quarantine action for Tagged Spam for monitoring purposes.

It's important that the *Incoming Spam* classification be located below those classifications that process email considered to be a greater threat to the environment. Generally, virus-infected emails have highest priority classification. Regardless of other characteristics, if an email contains a virus than this is to be the only classification assigned.

Classifications of other dangerous file types follow in the hierarchy. In our environment the detection of executables and encrypted emails are also ranked highly in the security policy.

The spam classifications were positioned in the hierarchy as follows:

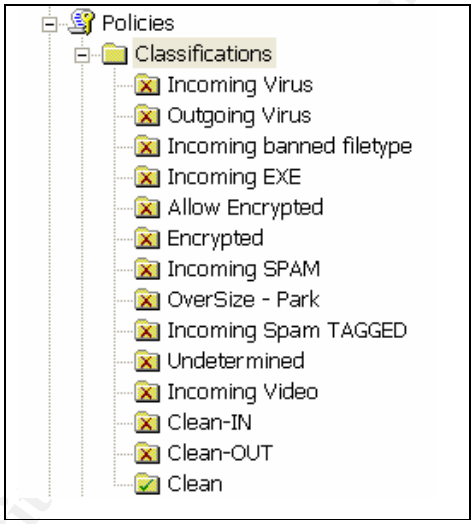

Figure 6: Classification Hierarchy with anti-spam filter.

2.3.2 Actions and Notifications

Having implemented the quarantine action for the spam classification, the next question became "should recipients be notified of quarantined spam?"

The advantages of including an action to notify the recipient include:
- Any genuine email would be recognised and released with minimum delay
- Recipients have control over what arrives in their Inbox

While the disadvantages include:
- Recipients have control over what arrives in their Inbox, which may circumvent the content security policy
- Bandwidth savings would be lost
- Users who receive large volumes of spam will now receive large volumes of spam notifications

I decided not to include a notification to recipients of emails quarantined as spam, and instead to focus on reducing false positives to be as close to zero as possible. As legitimate email that was tagged as maybe spam was still being delivered, my goal was to tune the filter so that only email that was almost certainly spam would be quarantined.

As I had chosen not to configure automatic notifications to the recipients of quarantined spam, it was my responsibility to see that any false positives were released as soon as possible. To this end, I regularly checked the quarantine folder to ensure that genuine emails were released every couple of hours. I also began saving copies of the wrongly-classified email into a folder that I would later use to create and update the not-spam reference file (NSASF). Once the NSASF was created, I could check the quarantine folders less frequently, but continued to do so at least daily. The risk of potentially losing business through lost emails warranted this initial investment of time, while the filter was being trained to my environment.

2.4 Updating the anti-spam filter

The spam reference file (ASF file) is available for download from a Clearswift site, and is updated eight times per day. The ASF file varies in size from 2-6 Mb, so I had to weigh the value of frequent updates with bandwidth and processing considerations. Due to the relatively small size of the environment I didn't think such a frequent update process was necessary, so I configured the update task to run once daily, at 5.00 am.

Once I had daily updates of the ASF file occurring automatically, I concentrated on creating a customised not-spam file (NSASF) to start reducing the false positives.

In order to do this, I had to save a large number of normal, clean emails that would become the base reference for what was not spam. A minimum of 5000 emails is recommended in order to create a comprehensive reference file.[11] Some of these emails I was saving from the spam and tagged spam quarantine folders, and the remainder I was gathering from the Clean-IN classification. It was important that all emails gathered be definitely not-spam from the corporate policy perspective.

I considered the two methods of collecting clean emails:

1. Implement a quarantine action (as well as deliver) within the *Clean-IN* classification. When stored in a quarantine folder, it is easy to view many emails at one time and quickly gather the truly clean messages.
2. Implement a save action (as well as deliver) within the *Clean-IN* classification. This saved a copy of the email to a designated folder. I then had to rename the file extension in order to use the Mailsweeper utility to view each email one at a time, and classify it as spam or not-spam. This utility then moved the email into a respective folder.

I found option 2 to be a much more time consuming means of sorting through the 5000 plus emails for classification. Although the utility is a useful tool in the classification of emails, it was overall too drawn-out a process for me to deal with at this time.

Option 1 provides a quick view of the sender, recipient and subject line of around 30 emails at a time (depends on screen resolution), without viewing the content of the email. I found that 99% of emails could be classified as spam or clean with just this information. The quarantine view also allows emails to be easily opened for further analysis if required.

[11] Clearswift Anti-spam Filter Installation & Configuration Guide, page 4-2.

06:29:56 PM 1...	Incoming Spam ...	hyigwsre@gmx.net	wdumpl More than 18 milion Americans suffer from some type of depressi
12:07:44 AM 2...	Incoming Spam ...	CHNZVV@ameritrade...	We can get you a refinance quote in 60 seconds.
08:41:54 AM 1...	Incoming Spam ...	CGJKPUBIKA@earthli...	We can get you a refinance quote in 60 seconds.
10:02:55 AM 2...	Incoming Spam ...	lnene8bcmjk@localac...	WE DELIVER UR NEEDED SOFTWARES TO ALL COUNTRIES AT CHEAP line
12:26:41 PM 1...	Incoming Spam ...	izg53cnu@icdc.com	WE GIVE THE BEST DISCOUNT EVER congestion drainpipe absorbate
03:29:12 AM 1...	Incoming Spam ...	dpxl5fcc@ttlc.net	WE GIVE THE BEST DISCOUNT EVER floor
04:24:47 PM 1...	Incoming Spam ...	vpnmcl82vqp@acces...	WE GIVE THE BEST DISCOUNT EVER many
05:49:31 PM 1...	Incoming Spam ...	nlh3je@qis.net	WE GIVE THE BEST DISCOUNT EVER supper nickel
06:09:36 PM 1...	Incoming Spam ...	acdlo31vros@vividne...	WE GIVE THE BEST DISCOUNT EVER taste
09:59:08 PM 1...	Incoming Spam ...	rtfewe62zy@tstonra...	WE GIVE THE BEST DISCOUNT EVER terrible key
06:36:09 AM 1...	Incoming Spam ...	jacksewellvc@wolfga...	We got what you need cheap! To your door! Overnight! ah
09:32:15 PM 1...	Incoming Spam ...	nux2fsq@tstonramp....	We Have 800 Expensive Softwares For U To Choose From $40 each incid
04:03:48 AM 1...	Incoming Spam ...	PKMSFAUKJTCIJM@g...	We have all prescription based muscle relaxants applique
04:04:10 AM 1...	Incoming Spam ...	D5PDPYU@dtl.co.nz	We have all prescription based muscle relaxants dunedin
07:12:21 PM 1...	Incoming Spam ...	cpfpvhqyvitsqj@netz...	We have the cheapest Viagra around!

Figure 7: Example view of Spam Quarantine folder.

Once I had collected the clean emails, I ran the 'asftrainer' program that is supplied with anti-spam filter. This program extracts the relevant information from the saved emails and pastes it into a not-spam text file, the NSASF file. Future emails can easily be merged into the existing NSASF file using the same program.

Each time that new data is merged into the NSASF, the Security Service must be restarted. This prompts integration of the NSASF data into the ASF file, and produces a valid, customised reference for the filter.

I used the same process to gather a sampling of spam emails that were being classified as clean. These I merged with the spam file (ASF) to produce a customised version that included both Clearswift's data and that from my environment.

So there it was. A fully customised anti-spam filter that would update itself with no further administration required. At last I could sit back, relax, and enjoy my spam-free environment.

3.0 What we have now (The "After")

Sit back and relax? Not quite. The new anti-spam filter is up and running but that's not to say there aren't ongoing issues that need to be addressed.

3.1 Post-implementation processes

Issue 1: Each time the updated spam text file (ASF file) is downloaded via the data feed, it overwrites the existing file. All training is therefore lost, including the spam file training and the not-spam file training. The new file contains the latest spam signatures, so I want to continue downloading this without losing the customisation.

Workaround: I edited the download script to include the 'asftrainer' merge function so that my collection of stored spam emails would be merged into the ASF file after every download. This adds only a few seconds to the download process, and my ASF file is now vendor-updated but still customised to my environment.

However, my new ASF file has still lost the not-spam data that had been previously combined from the NSASF file. In order to combine this information again, the Security Service needs to be restarted.

The following commands have been added at the end of the download script, to run after the new ASF file has been downloaded:

```
REM merge spam emails with new ASF file
asftrainer -s -d "d:\Saved_Spam" -m "c:\program files\mailsweeper for smtp\antispam\asf15.txt"

REM re-start security service so the NSASF data is imported into new ASF file
net stop "MAILsweeper for SMTP Security Service"
net start "MAILsweeper for SMTP Security Service"
```

Issue 2: Spam emails that should be blocked are either tagged as maybe spam or classified as clean, and in both cases delivered to users

Workaround: This is indicative of the need for an ongoing training process. Usually weekly, but more often if possible, I scan the Spam Tagged and Clean-IN quarantine folders and save the definite spam emails into the spam folder. These emails will be merged with the ASF file at the next download.

This has made it necessary to continue applying a quarantine action (as well as deliver) to Clean-IN emails. This was not part of the original plan but remains a useful means of quickly determining if any spam is getting through.

Issue 3: Legitimate emails continue to be tagged as Maybe Spam, particularly commercial purchase requests which contain similar phrases to spam emails. This is resulting in two issues:

3a. Users complain of [Maybe Spam] being appended to the subject line of legitimate emails. This occurs by design, to enable users to choose how they manage email tagged with [maybe spam] that arrives in their Inbox. It appears, however, that my environment is not suited to the implementation of this option.

Workaround: I edited the modification parameter in the filter configuration file (gas.xml) as follows: (edit in bold)

```
<classifications default="NOTSPAM">
        <classification>
            <threshold>5</threshold>
            <class>HIGH_SPAM</class>
            <modification></modification>
            <location>1</location>
        </classification>
        <classification>
            <threshold>1</threshold>
            <class>LOW_SPAM</class>
            <modification>[Maybe Spam]</modification>
            <location>0</location>
        </classification>
    </classifications>
```

Changing the <location> tag to 0 removes the [Maybe Spam] tag from the subject line, and instead inserts the following into the email header:

3b. Continued tuning of filter is required to prevent legitimate mails being tagged as possibly spam:

Workaround: As with the ongoing collection of spam emails for training the ASF file, so must non-spam emails continue to be collected and periodically merged into the not-spam text file. The Tagged Spam quarantine folder must be manually scanned for genuine mail which is then saved into a not-spam folder. Emails tagged as spam that are borderline are not merged into the not-spam file.

This ongoing training requirement means that, as with *Clean-IN*, the *Incoming Spam TAGGED* emails must to be quarantined as well as delivered. Due to the volume of email now being stored, the quarantine folders can only store up to three days worth before Mailsweeper performance suffers.

In addition to training, I have found it necessary to create a new Scenario Folder for emails originating from certain business partner and customer domains. The anti-spam filter is not active in this Scenario Folder. The commercial nature of emails from these domains results in a frequent classification of spam, which is not acceptable from a business perspective. Creating an additional "Non Spam Domains" list ensures that correspondence from these domains will never be classified as spam.

3.2 Impact analysis

The additional analysis of SMPT traffic has impacted on server performance, but not to the extent that corrective action is yet required.

There has been no real increase in email processing time since the anti-spam filter was implemented. Even with an increase in the volume of email over this time, average processing time has remained constant. There is, however, noticeable increase in CPU usage on the server. Where CPU used to average around 50-60%, it is now averaging at 80-100%. The more intense analysis performed by the Security Service explains this increase. The maintenance of the additional quarantine areas has also contributed to server load. This may be alleviated in the future by using the SAVE option instead of Quarantine.

The daily data feed updates combined with the spam and not-spam training, have proved to 'learn' quickly about new spam signatures. It is rare for spam not to be classified by its second day of circulation. I'm sure that increasing the frequency of the data feed may further improve this.

3.2.1 Spam detection rates

It has been four months since installation, and I continue to check the quarantine folders an average of twice a week. Over the last two months, only a few legitimate emails have been classified as spam, out of thousands. While legitimate emails are still tagged as maybe spam, they are mostly personal or borderline, and those that are business related are few in number. Even though tagging legitimate email as spam no longer has any impact on the recipient, periodic updating of the not-spam list will continue.

A Clearswift press release asserts that the Anti-spam filter has the following success rates in a customer proven situation:[12]

Spam Detection Rates: 92% success rate.
False Positives Rates: 1 in 1,000.

By analysing three days worth of email stored in the quarantine folders, I gathered the following statistics:

Clean IN:	3302*
Missed Spam (saved as clean):	125
Classified Spam:	1295
Tagged Spam:	490
All Spam:	1785
Of the Classified Spam:	8 were not spam**, and of these 4 were Type 4 commercial emails 1 was a joke-chain letter 3 were auto-generated notifications from a service provider
Of the tagged spam:	38 were not spam**, and of these 5 were business related 33 were personal, jokes, subscriptions, newsletters etc.
Total False Positives:	46

* This figure does not represent the total not-spam email as other classifications are triggered by attachment type, virus detection, sender address etc)

** As defined in the policy, ie: includes Type 4 and Type 5 email which should not be blocked.

With these statistics, I calculated the following:

Total Actual Spam: 1864
(Classified + Tagged + missed – false positives)

Actual Spam detection rate: 93.29%
(Classified + Tagged Spam – false positives / Total Actual Spam * 100)

Spam quarantine rate: 69.05%
(Classified – false positives (classified) / Total Actual Spam * 100)

False Positives[#]: 0.61%, or 6.1 per 1,000
(Classified false positives / Classified Spam)
[#] Mail quarantined as spam, causing delay or loss.[13]

Business emails trapped as False Positives: 0.54% or 5.4 per 1,000
(Classified business false positives / Classified Spam)

Overall, these figures correspond with Clearswift's assertions. The false positive rate is a little high, however with further tuning I expect to reduce this. The emails that were

[12] Clearswift Press Release
[13] Snyder, Joel.

falsely quarantined may have been important circulations so this is a concern. I expect that by adding these to the not-spam reference file lower false positives will result, with the likely side effect of slightly increasing the amount of spam accepted as clean. This is what I believe will be an on-going process, to continually adapt to the changing characteristics of both genuine and spam email.

In addition to these statistics, I took sample figures from random weekdays both before and after the filter was implemented. I found that before the new filter was installed, approximately 3.76% of email was being classified as spam, and now the average is 22.24%. Some days can produce spam rates as high as 43.90% (see Appendix 1 for details).

3.3 Future considerations

One of the main considerations for the future is whether or not to bring users into the training process. If users have the power to determine what is spam and what isn't, it will reduce the load on the Administrator, as well as solve the problem of genuine email being lost in quarantine. However, individual user perception may contradict the corporate security policy, so this is not considered an option as this time.

Another alternative may be the introduction of desktop-based spam filters configured individually by each user. Each mail recipient would then have individual control on what happens to email that is addressed to them. While this would give users more control over their correspondence, it may potentially cause breaches in security policy if pornographic or offensive email is distributed internally.

Taking control of policy implementation out of the Administrator's hands is, therefore, not an option at this time.

The future education of users in email and internet protocol will continue, specifically:

- Work email addresses should not be used for subscribing to anything other than business-related activities
- Only subscribe to sites that promote a clear unsubscribe procedure
- Only subscribe to sites that have a published privacy policy guaranteeing that email addresses will not be given to third parties
- Spam emails should not be replied to, even in an attempt to unsubscribe. These emails are to be forwarded to the Administrator for their inclusion in the ASF file
- Continue to advise the Administrator of important subscriptions etc, so these can be added to the NSASF or non-spam domain list as soon as possible.

4.0 Conclusions

Spammers will continue to spam as long as there are acceptable returns; as long as people continue to buy the products.[14] Defeating spam therefore depends both on user education and the development of technology that learns very quickly to adapt to new spamming techniques.

Many email users consider spam to be their greatest problem. That is, until they stop receiving desirable email that is trapped in an anti-spam filter. The toughest aspect of protecting an environment from spam is finding the balance between how much spam you are willing to accept for the assurance that no genuine email will be lost.

[14] Hurley, Edward (quoting Ron Scelson)

There is no single solution to spam. Technology implemented at the gateway is an important contributor to an anti-spam policy, but is just one part of the solution. Whichever technologies are implemented, they must be dynamic and adaptable to new techniques or risk becoming outdated in this ever-changing environment.

References

Hurley, Edward. "In the spammer's lair". SearchSecurity.com News. 16 Jul 2003. URL: http://searchsecurity.techtarget.com/originalContent/0,289142,sid14_gci914837,00.html (June 2004)

Levitt, Mark & Burke, Brian E. "Choosing the Best Technology to Fight Spam". IDC White Paper. April 2004.

Leyden, John. "Phatbot arrest throws open trade in zombie PCs". The Register. 12 May 2004. URL: http://www.theregister.co.uk/2004/05/12/phatbot_zombie_trade/ (June 2004).

Lyman, Jay. "Xombe Trojan Spoofs Microsoft Patch To Steal Personal Info". TechNewsWorld. 12 January 2004.
URL: http://www.technewsworld.com/story/32573.html (June 2004).

Snyder, Joel. "Spam and Statistics". Network World Article. Network World Fusion. 09/15/03. URL: http://www.nwfusion.com/reviews/2003/0915spamstats.html (June 2004).

AUSCERT ALERT AL-2004.10. "Bogus Banking Email Allows Trojan Infection for Outlook Users". 04 April 2004. URL: http://www.auscert.org.au/render.html?it=3981 (June 2004).

Clearswift Ltd. "Clearswift adds First Self-learning Anti-spam Solution to CS MIMEsweeper™ Portfolio". Clearswift Press. 12 November 2003.
URL: http://www.clearswift.com/news/PressReleases/225.aspx?pr=true (June 2004)

Clearswift White Paper. "Spam: Corporate Concern, Personal Perspective". Clearswift Ltd. February 2004.

Clearswift Press. "Anti-Spam Filter Installation and Configuration Guide". Clearswift Ltd. Rev 1.0, November 2003.

Clearswift Press. "Clearswift Mailsweeper for SMTP Reference Guide". Clearswift Ltd. Rev 2.2a, August 2002.

LURHQ Threat Intelligence Group "Sobig.a and the Spam You Received Today". 21 April 2003. URL: http://www.lurhq.com/sobig.html (June 2004).

Net Sense. "Spam Solutions White Paper"
URL: http://www.netsense.info/Spam%20Solutions%20White%20Paper.pdf (June 2004)

Author's Note:

This paper has not been reviewed or in any way endorsed by Clearswift Ltd. Any comments on Clearswift products are the sole opinion of the writer.

Appendix 1: Amount of spam detected as a % of all email

Results from random sampling of detected spam, before and after Anti-Spam filter implementation.

	Previously	With Anti-spam filter						
Sample Date	Detected Spam	Classified Spam	Tagged Spam	Emails	% Classified Spam	% Classified and Tagged Spam	Ave % Classified Spam	Average % Classified and Tagged Spam
Sept 2003 (Mon)	148			4547	3.25			
Oct 2003 (Tue)	163			3503	4.65		3.76	
Nov 2003 (Wed)	143			3662	3.90			
Dec 2003 (Thu)	150			4654	3.22			
Jan 2004 (Fri)		475	651	3893	12.20	28.92	12.79	22.24
Feb 2004 (Mon)		484	660	5044	9.60	22.68		
Mar 2004 (Tue)		446	336	4019	11.10	19.45		
Apr 2004 (Wed)		627	255	4241	14.78	20.80		
May 2004 (Fri)		602	209	3999	15.05	20.28		
Jun 2004 (Thu)		575	300	4106	14.00	21.31		
May 2004 (Sun)		481	142	1419	33.90	43.90		