



blackspider
technologies

Spam: now a corporate concern



Whitepaper Contents

Introduction	2
What is spam?	3
The Cost to Business.....	4
Direct IT Costs.....	4
Productivity Costs.....	4
Who are spammers and how do they get my address?	5
What techniques exist for detecting spam?.....	6
MailControl Spam	8

Introduction

You have probably already come across the problem of junk e-mail, commonly known as spam. You may have received e-mails trying to sell you cheap Viagra, free credit cards, university degrees and subscriptions for various 'adult' websites.

Over the last 18 months the escalation in spam has been alarming, with statistics from Internet Service Providers such as Yahoo and AOL showing that over 50% of all e-mail they process is junk.

What was once a mild irritation has now grown into a global epidemic, with companies beginning to realise that spam is becoming a serious drain on resources and a threat to e-mail security. This white paper discusses the problem of spam, where it comes from, the impacts on companies and how best to protect businesses from this growing menace.



What is spam?

Spam is the brand name of a luncheon meat made by Hormel Foods, but it is also the name used on the Internet to describe junk e-mail.

Definitions vary, but a number of characteristics of a message define it as spam.

- The e-mail message is sent to a large number of people advertising a product or service.
- The recipients of the message did not request (and typically don't want) to receive advertisements from the sender.
- The sender of spam frequently attempts to hide or obscure their identity.
- The advertiser will continue to send unwanted e-mails even when requested to stop by the recipients.

Various stories explain how this type of e-mail got the name spam. The most likely is that spam gets its name from a television comedy sketch by British comedians Monty Python. In the sketch, a restaurant serves Spam (the luncheon meat), with any order placed. Customers could order "Spam, Spam with a side of Spam", and so on. An annoyingly funny waitress repeats the name "Spam", over and over again. This is accompanied by a number of Vikings singing a "Spam song" louder and louder until they drown out the whole restaurant.

The point? Spam is something that you get whether you order it or not and eventually the noise of 'spam' will drown out everything else.

The volume of spam really is beginning to drown out valid e-mail. In June 2003 over 57% of all e-mail processed by BlackSpider Technologies was classified as spam. In one extreme case a customer with less than 10 employees was receiving over 3,000 spam messages a day making their e-mail system almost unusable.

European Commission figures estimate that spam cost businesses \$10 billion globally in 2001 and since that time spam volumes have more than doubled. Despite action from the US Government and the European Commission the tide of spam is only set to increase.



The Cost to Business

Direct IT Costs

Spam has some obvious direct costs. Unlike standard postal mailing where the sender pays the postage, the recipient of spam has to pay to receive the e-mail.

Delivering the message costs money in Internet bandwidth and disk storage. Once received the unwanted message has to be stored until read and deleted by the user. In many cases the e-mail still remains in the 'deleted Items' folder even after it has been discarded.

Although the cost of a single spam message to an individual user is small, the aggregate cost to an organisation can be considerable.

Productivity Costs

Along with the direct costs associated with IT resources there is a large indirect cost in terms of employee productivity. Apart from being very irritating, managing and deleting unwanted e-mail takes time and effort. Many spammers are using techniques to disguise e-mail and to trick users into opening messages. As the quantity of spam increases, so does the cost involved in managing it.

Let us assume that opening and deleting an unwanted e-mail takes 10 seconds, and each user receives on average 4 each day. By extrapolating those number for a company with 500 users (40 seconds * 20 days * 12 months * 500 users / 8 hour day), 166 man days will be lost each year, just dealing with spam.

Business Risk

Spam is unregulated. In most countries bulk mailing by post is regulated by a government organisation, but due to the nature of the Internet, spam is not. This means there is no control over the content of e-mail. Many unsolicited messages contain pornographic and sometimes deceitful advertisements, such as the infamous 'Nigeria 419' scam offering a share of millions of dollars in exchange for a 'small' upfront administration fee.

Pornographic and offensive spam create a specific problem for businesses, because employers have a 'duty of care' to protect their staff from these types of messages. There have already been well-publicised cases where employees have successfully brought legal cases against employers, primarily in the area of sexual harassment via e-mail. Considering the growth in pornographic spam and the increasingly litigious business environment, the legal risks associated with not protecting employees must be considered.



Who are Spammers and how do they get my address?

By their very nature spammers are difficult to profile. However, it is clear they have access to significant IT resources and believe they can make a profit out of this form of bulk advertising.

Most spammers are selling a product or a service and often go to great lengths to disguise their identity. In many cases their e-mail message will contain a forged header, occasionally looking like you sent it to yourself, in an attempt to get you to open the message.

Much of the world's spam emanates from the USA, where one million e-mail addresses can be purchased for as little as \$1 US dollar. It's easy to understand that even with very small responses to their spam advertising; Spammers are still able to generate significant profits from their activities.

Spammers have developed a number of different techniques to harvest e-mail addresses. Originally spammers targeted news groups and bulletin boards, writing software to trawl through and capture e-mail addresses.

Increasingly spammers are directly targeting corporate networks using a technique called a 'directory harvest attack'. During an attack spammers attempt to deliver messages to multiple addresses, such as bill.smith@acme.com, bsmith@acme.com, and bill@acme.com. Addresses that are not rejected by the receiving mail server are determined as valid. These addresses are compiled and sold to other spammers worldwide. Within hours, a brand new email box can be full of unsolicited, junk email.

In addition spammers are harvesting e-mail addresses directly from company websites and targeting e-mail addresses they know will be read, such as info@acme.com or support@acme.com

As well as harvesting e-mail addresses, spammers have developed techniques to see if their messages have been opened, thereby validating the e-mail address.

By embedding HTML code inside the body of an e-mail message with links back to a website, spammers know when, and who opened their messages. These links are triggered when the e-mail is opened either by clicking on the message, or in the case of many e-mail clients when the message appears in the 'preview' panel.

Often at the bottom of a spam message you will find some 'unsubscribe' instructions or a link. Ironically, these are designed to tempt you into responding, again validating the e-mail address while the spammers have no intention of removing your details from their lists.

Combinations of these techniques mean that spammers continue to net thousands of new corporate email addresses. The results force unprotected corporations to incur higher email system costs, face increased breaches in security and decrease their email system's reliability.



What techniques exist for detecting spam?

Detecting and filtering spam creates a number of complex challenges due to the dynamic nature of junk e-mail. An effective spam filter must block the maximum unwanted e-mail, with the minimum number of false positives (messages, wrongly identified as spam).

This is further complicated by the fact individual users have different views on what they consider to be spam. Many users are very happy to receive adverts from well-known Internet retailers while others consider this junk.

There are a number of techniques available today for spam detection, including:

Real Time Black Lists: These are collaborative services operated on the Internet by commercial organisations, or communities of interested users. RBLs contain lists of IP addresses of machines that have been black-listed for various reasons. Machines are black-listed because they may have been used to send junk e-mail before, or they are configured as 'open relays' meaning they are not secure and can be hijacked by spammers.

RBLs also include lists of IP addresses associated with ISPs dial-up user accounts. These are accounts used by home users that typically would not be sending e-mail directly, but would be routing e-mail through their ISPs mail servers.

RBLs are a useful tool and can be used as an indicator that a message may be spam, but should not be used as an absolute guarantee as machines are often incorrectly black-listed and with some RBLs, it is notoriously difficult to get removed from the list.

Lexical Analysis: is a term used to describe the analysis of an e-mail message looking for indicators that the message may be spam, or valid e-mail. The concept of lexical analysis started out looking for key words and phrases inside the body of a message for strings that would be commonly found in spam, such as 'Buy Cheap' and 'Get Rich Quick'.

More sophisticated lexical analysis engines now look at the whole message, including the message envelope, the message headers, the subject line and the body text. There are often key indicators in the message header that would suggest the message is spam. For example, having no 'from:' in the message envelope or having the same e-mail addresses on the 'to:' and 'from:' line of the message header are often indicative of spam e-mail.

As well as looking for spam, it is possible to check for characteristics found commonly in valid e-mail, for example to the message ID and message headers indicate that the e-mail was sent from a Microsoft Exchange server.

While no one test is sufficient to positively identify a spam or valid e-mail, by taking a heuristic approach to lexical analysis it is possible to improve the spam detection rates.

Distributed Checksum Clearinghouse: DCC is a client/server system developed by an Internet community lead by Vernon Schryver. This collaborative system works on the basis that servers using DCC create a 'fuzzy hash' for every e-mail processed. The fuzzy hashing logic allows e-mails to be compared, identifying similar messages that might contain slight variances, such as different greetings.

These hashes are then collated at the Clearinghouse and counted. The highest scoring e-mails are the ones that have been seen most frequently across the Internet, and are likely to be spam.

Bayesian Filtering: Using Bayesian inferential statistics is a relatively new innovation in spam detection. The concept of Bayesian filtering is to create two databases or 'corpuses' of e-mail: A corpus of spam e-mail and a second of valid e-mail.



Each corpus is then 'tokenised' and analysed looking for tokens that frequently appear in each type of e-mail. Each token is then given a probability weighting, suggesting if it is likely to appear in spam or valid e-mail.

Each new message is then processed and tokenised, and the tokens compared against the existing database to determine the probability that the message is valid or spam.

This approach creates some unusual but effective results, for example the token: "ff0000" frequently appears in spam corpuses. This is the HTML tag for bright red, something spammers often use to highlight their special offers.

A key benefit of this approach is that it is possible to tune the filters to different customer environments. For example the token 'Viagra' would typically appear in the corpus of Spam with a high Spam probability. However for pharmaceutical companies selling Viagra it is also likely to appear in their valid e-mail; in this case 'Viagra' would be a less significant indicator of spam.

White & Black Lists: As well as online real-time black lists, there is a place for user configured white and black lists. These are configurable lists of e-mail addresses (or domains) that organisations explicitly block or allow through. These are particularly useful to tune spam detection systems for individual 'corner case' scenarios.

MailControl Spam

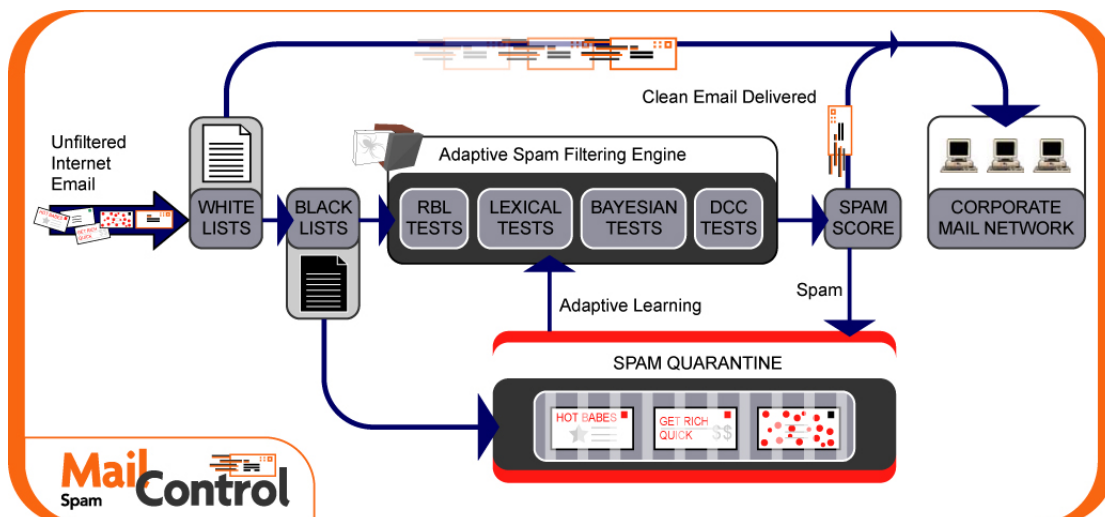
MailControl Spam from BlackSpider Technologies is a fully managed service designed to provide companies with the highest level of protection available at a cost effective price.

At BlackSpider we believe that no one approach to Spam detection is a silver bullet, only by combining the best spam detection techniques, with a world-class infrastructure and industry expertise can you deliver an effective solution against the growing menace of spam.

The strength of the MailControl service lies in its ability to combine the power of lexical analysis, real-time black lists, Bayesian probability, customer control white and black lists, and collaborative hashing techniques into an adaptive filtering engine, which provides the highest levels of spam detection and almost eliminates false positives.

This approach combined with the ability to set spam thresholds on a per-user or domain basis ensures that MailControl Spam is the most effective and usable technology on the market today.

The whole process takes less than one second. Once each message has been analysed the message receives an overall 'spam score'. The score is then compared against the spam threshold defined by the customer; mail scoring below the threshold is delivered as normal, whilst mail scoring above is classified as spam.



A key factor in any successful anti-spam solution is creating and retaining user confidence in the service. MailControl achieves this by giving end-users visibility of the spam detected by the service without creating a volume of unwanted e-mail. Users can request an individual report from MailControl, providing details of all their e-mail processed with the 'spam score' of each message.

This approach ensures users retain confidence in the service, and that their e-mail is being correctly classified.

In summary, MailControl Spam controls the flood of unwanted e-mail at the Internet level, keeping it away from your networks, increasing employee productivity and reducing costs.

Please do not hesitate to contact us if you need further help or assistance:

Sales Desk: +44 (0) 1189 653 700 E-mail: sales@blackspider.com

Operations Desk: + 44 (0) 1189 653 800