

INTRODUCTION TO USING SPAM METHODOLOGY TO INITIATE PROACTIVE SPAM CONTROLS

Nithen Naidoo

SensePost

nithen@sensepost.com

(012)667 4737

P.O Box 10692, Centurion, 0046

South Africa

ABSTRACT

The paper discusses the multi-faceted evolutionary nature of spam on the Internet and the efficiency of current control mechanisms.

Since many people are unaware of how large or lucrative the spam industry actually is, in the first section the author will discuss the magnitude of the market and key players in the spam arena. This will enable readers to understand why spammers will consistently go to extreme lengths to bypass current spam filtering technology.

In the second section “Mail made simple” the author will briefly describe how email technology works and integrates into the network environment. Basic email forensic techniques will also be discussed.

Spam by nature is very dynamic and often difficult to define. Without a precise definition and classification system identifying spam email accurately will not be possible. In the section “Identifying and classifying spam”, the author will define numerous spam signatures and behaviour patterns.

The section entitled “Spamming techniques, tools and methodology” describes underground tools and data mining techniques used by spammers. The author also looks at driving forces behind spam technology development.

“Spam evolution” tries to establish exactly how spam evolves and thus manages to stay one step ahead of current security technology.

“Control mechanisms”, the final section, discusses numerous proactive control mechanisms and techniques used to combat spam. This section will use knowledge gained by the research information to develop effective and efficient counter measures.

KEY WORDS

Spam, Email, Fraud, Security

INTRODUCTION TO USING SPAM METHODOLOGY TO INITIATE PROACTIVE SPAM CONTROLS

1 INTRODUCTION

The Internet has introduced the world to a façade of electronic freedom. This “freedom” is often exploited by malicious or opportunistic members of the Internet community. A prime example of such an injustice is spam. There are numerous definitions for spam.

MAPS (Mail Abuse Prevention System) www.mailabuse.org defines an electronic message as spam if: *“(1) The recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients; AND (2) the recipient has not verifiably granted deliberate, explicit, and still-revocable permission for it to be sent; AND (3) the transmission and reception of the message appears to the recipient to give a disproportionate benefit to the sender.”*

Many people believe spam commonly refers to UCE (unsolicited commercial email). But UCE is merely a small facet of what spam represents today. In actuality most spam is not UCE. Dr Curtis Kret of the Secure Science Corporation (www.securescience.net) defines spam simply as “*undesirable email*” falling into one of the following common classes:

- UCE
- NCE (non-responsive commercial email)
- List makers
- Scams
- Covert messages camouflaged as spam

Due to the dynamic nature of the Internet and its rapid development these definitions are constantly re-assessed. Many corporations now have included a class of spam described by research firm Gartner as “*friendly fire*”. These emails are usually sent by family and friends with large attachments which hog bandwidth and reduce employee productivity.

Spam much like the Internet is constantly and dynamically evolving. It is multi-faceted and therefore defies precise definition. The only certainty is that it is a growing problem facing the Internet community at large, so much so that governments and large corporations (e.g. Microsoft) have made spam a primary concern for 2004. Eric Allman, the creator of the world’s first Internet mail program, says “There is a genuine concern that too much spam will kill off email”.

2 WHY SPAM? – SHOW ME THE MONEY

This section will try to establish how large and lucrative the spam industry actually is. The researcher will identify key players, their roles and their source of motivation. The Spam industry has grown at a phenomenal rate and therefore it is difficult to define key roles and market figures. Current research suggests that the key players in the commercial spam industry fall into three major categories namely list makers, bulk mailers (spammers) and service buyers. A fourth category - scam mailers - often falls between the bulk mailer and service buyer categories depending on their skill level and technical know how.

List makers are the individuals who collect email addresses. They use various techniques to collect email data and sell different grades of addresses based on the validity of the data. Most of the list makers encountered during the research project were between the ages of 16 to 24 and often

operated from school and university networks. The average price per CD of email addresses sold is roughly \$100 depending on the grade of data. Although they perform a key role they certainly are not the big earners in the market space.

Bulk mailers or Spammers are an entirely different breed. During April 2004 Scott Richter founder of optinrealbig.com was being sued by the State of New York's Attorney General for violating federal deceptive marketing laws, including misleading subject lines and faked senders' addresses. Essentially Richter is being accused of spamming or "email marketing" as he calls it. Richter's optinrealbig.com is the third largest advertisement mailer on the net according to the law suit claims. He has also been listed as one of the "ten most influential and powerful men under 38" by Details magazine. Spammers or bulk mailers seem to reap far greater rewards for their services. This would suggest that they are the big earners in the spam market arena.

Spammers are capable of bulk mailing more than 250 million email messages a day. Although the market prices may vary optinrealbig.com charges around \$200 for a million email messages sent. Anti-Spam company MessageLabs say that 80% of mail sent worldwide is attributed to Spam and costs UK businesses around £3.2 billion a year. The market is so large that the Spam kingpins now have their own ISPs and run Spam networks. Spam networks allow spammers to send thousands of messages from hundreds of different points around the world. This makes the big players in the Spam industry difficult to trace and convict.

It is difficult to predict how much the service buyer or scammer gains from the spam epidemic (values may vary). The only good indication at this point is that they are willing to pay \$200 per million messages sent and that they are responsible for more than 65% (marketing and scam mails collectively) of mail sent worldwide. This would suggest that the spam industry is certainly large and extremely lucrative.

3 MAIL MADE SIMPLE

The reader will need a basic knowledge of email architecture and mail headers to truly comprehend the sections that follow. This section tries to familiarize the user with a few basic mail concepts.

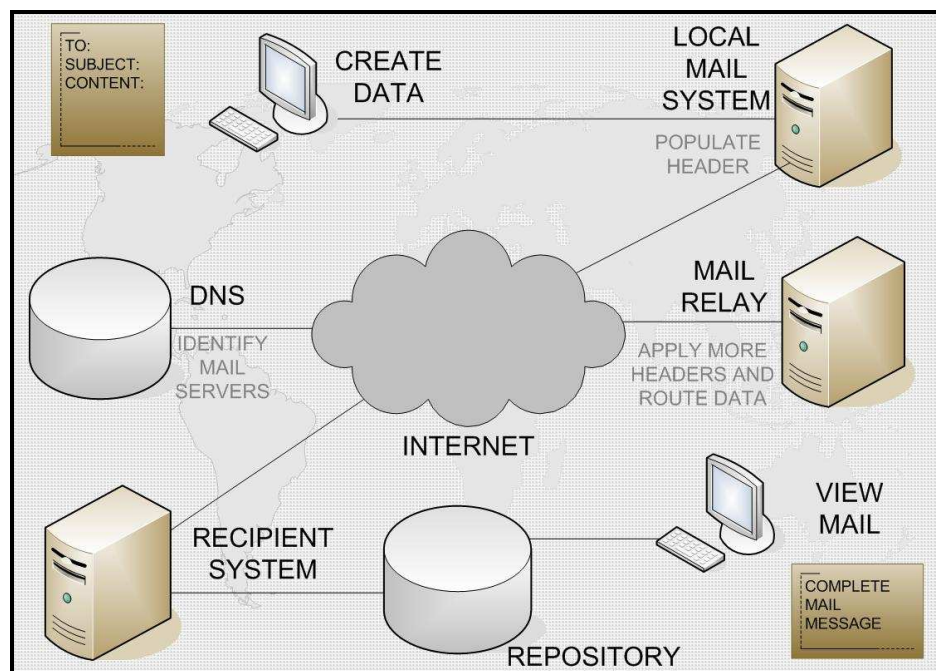


Figure 1 - Email delivery diagram

After a message is created and it leaves the client it contains the “To”, “From” ”Subject” and “Content” fields. These fields are described in *Table 1* and can be seen in the email header example in *Figure 2*. There are numerous other fields (“date and time”, “CC” and “BCC”) but we are only concerned with the basic format for now.

Table1 - Basic field inputs to an email

Field	Description
To	To whom the message is to be sent
From	From whom the message is sent
Subject	Brief description given by the creator
Content	Actual mail message

```

Received: from rauteg.rau.ac.za [rauteg.rau.ac.za [152.106.1.53]]
Iby GrasGroen.sensepost.com [8.12.10/8.12.7] with ESMTP id i319smBk002790
Ifor <nithen@sensepost.com>; Thu, 1 Apr 2004 11:54:49 +0200 [SAST]
Received: from frodo.rau.ac.za ([152.106.2.140] helo=frodo.NetworkAl.local)
Iby rauteg.rau.ac.za with smtp [Exim 4.22]
Iid 1B8yXw-0000ew-9y
Ifor nithen@sensepost.com; Thu, 01 Apr 2004 11:31:36 +0200
Received: From II ([152.106.42.233]) by frodo.NetworkAl.local [WebShield SMTP v4.5];
Iid 108081186031; Thu, 1 Apr 2004 11:31:00 +0200
From: "Prof Les Labuschagne" <LL@na.rau.ac.za>
To: "nithen" <nithen@sensepost.com>
Subject: RE: ISSA2004-Absract Submission
Date: Thu, 1 Apr 2004 11:31:36 +0200
Message-ID: <004301c417cc$20be82e0$e92a6a98@II>
MIME-Version: 1.0
Content-Type: text/html

```

Figure 2 – Example of an email following the RFC-2822 format

After the message is received by the local mail server it is given an initial header (received by), this header appears as follows:

- Received: from [sending-host's-name] [sending-host's-address]
- by [receiving-host's-name]
- [software-used]
- with [message-ID]
- for [recipient's-address]; [date][time][time-zone-offset]

Two examples of such headers can be seen in *Figure 1* in red and green text. The message then progresses through numerous Mail Relays where the message is given appended header information. The mail is eventually received by the recipient’s mail server and stored in the recipient’ mail account (Inbox) where it is downloaded by the user. At this stage the message has received a final header. Additional information given by the headers include Message IDs, MIME Version and Content type

MIME (Multipurpose Internet Mail Extensions) is a standard for handling various types of data. This essentially allows you to view mail as either text or html. There are other MIME types defined which enable mail to carry numerous attachment types. A Message ID is assigned to a

transaction by a particular host (the receiving host or the 'by' host). These message IDs are used by administrators to track transactions in the mail server logs.

Most spammers forge email headers, this makes spam email difficult to trace back to the original source. There are a few fields (as shown in the previous section) that can be forged. These are as follows:

- Recipients: To / From / CC
- Subject / Date / Message / ID / Content
- Initial headers

The following fields cannot be forged but may be misleading when analyzed.

- Time stamps – Time stamps are dependant on the local host's time clock which may be inaccurate.
- Originator IP Address – There are numerous techniques used by spammers to send or relay mail from compromised hosts on the Internet (these techniques will be discussed later in the paper). Although the data may be accurate it may not necessarily lead back to the offending spammer.

4 IDENTIFYING AND CLASSIFYING SPAM USING MAIL HEADERS

Spam has a versatile evolutionary nature, which makes it difficult to define and classify. The market continually finds new and innovative ways to bypass filters, evade authorities and eventually sell a service or product. It is essential to have an accurate definition and classification system if the Internet community wishes to rid themselves of this epidemic. For identification to be truly effective and efficient mail analysts need to look at the most static features of a spam email. The most static properties of a single spam mail would be found in the mail headers.

For example the US FTC (Federal Trade Commission) claims that in 2003 33% of all spam mails had false "From" headers. In actuality most spammers forge headers and these mismatches in header recreation can be used as static signatures for filtering.

A good example is a list maker's spam mail. A list maker collects email addresses and grades them as follows:

- Grade A – user exists and email account is in use
- Grade B - user exists but the account may not be in use.
- Grade C – user does not exist.

Typical list maker signatures would included forged headers, MIME type HTML for "Web-Bugs" (discussed later in this paper), valid URLs, a unique recipient and most importantly the "From" and "Reply-To" fields will match. List makers require some type of response from their email, which is why certain fields are valid ("Reply-To" and "From"). List makers need to keep state of the addresses that the actual response was from therefore they have unique recipients. With other forms of spam most headers are forged and have multiple recipients.

As can be seen by in the example above, there are numerous static signatures in spam mail headers. Not only will this help analysts categorize spam, it can help filters accurately tag spam mail. Header information contains far more accurate signatures for filter rule development. These rules are less likely to be bypassed as compared to content filtering technology.

5 SPAMMING TECHNIQUES, TOOLS AND METHODOLOGY

Spammers often use custom tools that they develop themselves to send mail, bypass filters, collect data and hide their true identities. This contributes largely to Spam's evolutionary nature. During this research project many commercial tools were discovered that offer the same functionality. If the tools spammers use are studied, much in the same fashion as a spammer probably studies a spam filter, it may be possible to find weaknesses in the spammer methodology. Spammers are after all human and often act in a very predictable manner. In most cases spammers follow a strict methodology and only change if the need arises. This section will describe spammer methodology, numerous tools and spammer techniques. Tools used by spammers and list-makers fall into 2 classes:

- Email address harvesting
- Bulk mailers

This may seem like a simple classification system but within these two classes breed tools of vast diversity. The tools encountered during the research project resembled those used by professional ethical hackers and virus writers alike. Many industry specialists view numerous Trojans as specialized spammer tools.

5.1 Email address harvesting

The 3 most commonly used tools for mining email addresses are:

- Email extractors (robots)
- Email spiders
- Email brute forcers

These tools are used by list makers to mine email data. The most commonly used technique is email extraction, which is performed by software programs called 'robots'. These automated extractors browse Internet pages and extract email addresses directly from web content. The Centre for Democracy and Technology (2003) rated public web posted addresses as the biggest source for spam address collectors.



```
X-Win32
[server@infosecsa.co.za]
> perl extract-mail.pl www.infosecsa.co.za/contact.html

Checking www.infosecsa.co.za/contact.html - try 1..
Checking www.infosecsa.co.za/contact.html
[amanda.ce@up.ac.za]
[eloff@cs.up.ac.za]
[info@cyansky.co.za]
[martha.ce@up.ac.za]
[l1@na.rau.ac.za]
[eloffmm@unisa.ac.za]
>
>
```

Figure 2 – Example output of a simple email extractor

Email extractors are very effective and widely used, but they harvest data without discrimination. Email 'brute forcers' and 'spiders' seem to be developed by list makers targeting a more select market space. The list maker uses a domain name (e.g. sensepost.com) as a keyword in a simple search queries, he then mines data from numerous web search engines (e.g. google.com). The process can be easily automated, and aids the spammer who is interested in a specific market space.

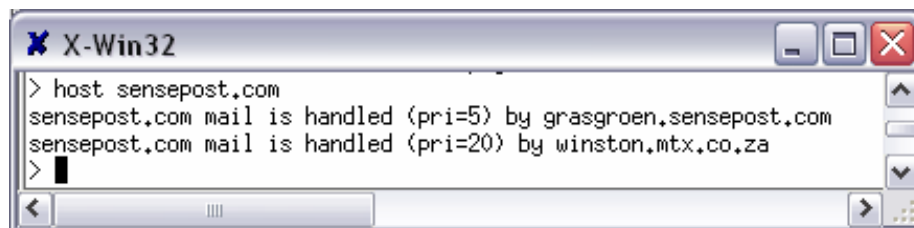


```
> perl emails.pl
emails.pl domain mode
Eg. emails.pl sensepost.com 1 (1-emails, 2-sub domains)
> perl emails.pl sensepost.com 1

info@sensepost.com
haroon@sensepost.com
charl@sensepost.com
-roelof@sensepost.com
drubin@sensepost.com
research@sensepost.com
roelof@sensepost.com
christoff@sensepost.com
>
>
```

Figure 3 – Example output of a simple email spider

The spammer or list maker could also opt to brute force the mail server. A simple host lookup reveals the primary mail server for the specific domain.



```
> host sensepost.com
sensepost.com mail is handled (pri=5) by grasgroen.sensepost.com
sensepost.com mail is handled (pri=20) by winston.mtx.co.za
>
```

Figure 4 – Example of a host lookup

Thereafter the spammer or list maker brute forces common user names against the domain to verify valid users.



```
> telnet grasgroen.sensepost.com 25
Trying 196.30.67.6...
Connected to grasgroen.sensepost.com.
Escape character is '^]'.
helo grasgroen.sensepost.com
220 GrasGroen,sensepost.com ESMTP It's patched...Gaan weg julle kube
Fri, 30 Apr 2004 15:05:18 +0200 (SAST)
250 GrasGroen,sensepost.com Hello [65.61.162.77], pleased to meet yo
mail from;nithen@sensepost.com
250 2.1.0 nithen@sensepost.com... Sender ok
rcpt to: alicen@sensepost.com
550 5.1.1 alicen@sensepost.com... User unknown
rcpt to: ben@sensepost.com
550 5.1.1 ben@sensepost.com... User unknown
rcpt to: charl@sensepost.com
250 2.1.5 charl@sensepost.com... Recipient ok
rcpt to: david@sensepost.com
550 5.1.1 david@sensepost.com... User unknown
```

Figure 4 – Example of a brute force on a mail server

List makers often use far simpler techniques to mine additional valid email addresses. The chain letter is an excellent example. Would the phrase “Really it is an excellent example, It’s has been on the news even on Oprah”, ring any bells? The list maker gains the unsuspecting victim’s trust either using sympathy or a promise of rewards. The email user is asked to add his friends’ email addresses, forward the mail to 10 people or pass along the petition to save the endangered pink elephant. The unsuspecting recipient does not realize that he is actually aiding the list maker to mine valid email addresses.

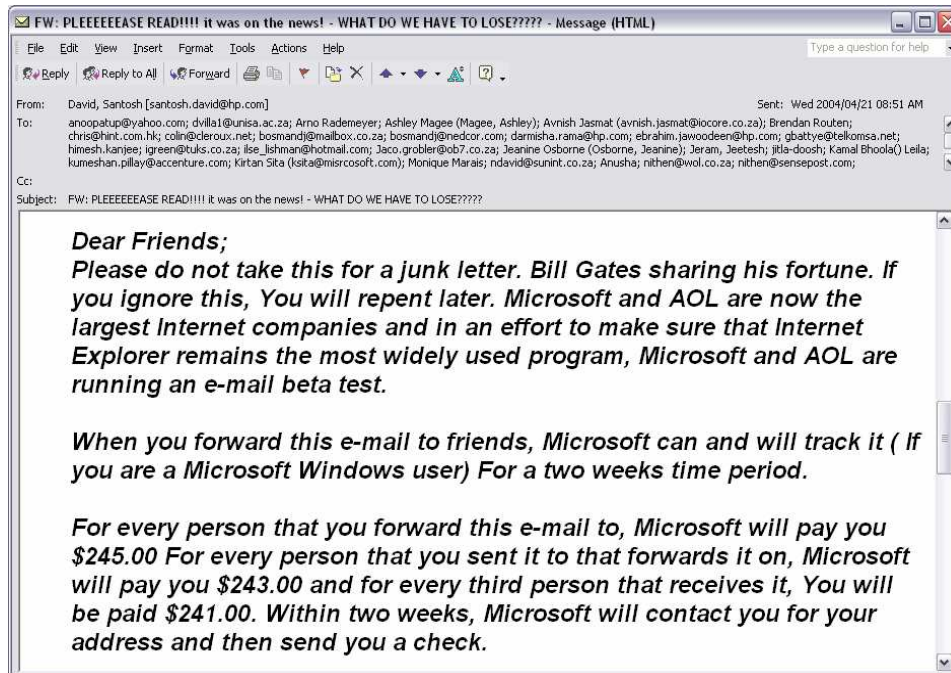


Figure 4 – Example of a chain letter

5.2 Bulk mailers

There are numerous bulk mailing tools sold on the market today, the cutting edge spammer technology not only allows for bulk mailing but helps hide the spammer’s true identity. Earlier in the paper the researcher mentioned techniques used by spammers to conceal their IP address from the mail headers. This section will discuss three of these techniques namely:

- Mail transfer via open relays
- Web site mail-form hijacking
- Mail via open proxy (compromised hosts)

Most ISPs will not tolerate a spammer’s traffic load, disgruntled subscribers and spam victims often report spam mail back to the originating ISP. Spammers are forced to either to find spam friendly ISPs, which are now rare in developed countries, or to transfer mail through hosts allowing indiscriminate relaying. The second option mentioned is commonly referred to as transfers via open relays.

Web applications have become a key focus area in Internet security. Spammers exploit insecure mail scripts which allow them to send bulk emails using an unsuspecting victim’s mail service. This not only allows them to bulk mail without ISP interference but also allows them to remain anonymous. This particular form of mail-form hijacking occurs when the senders address is posted as a hidden field. The spammer intercepts the request and alters the post data. The senders address is changed to the spam victim’s address and then re-posted. The following two figures demonstrate this exact mail-form vulnerability on a random web mail form.

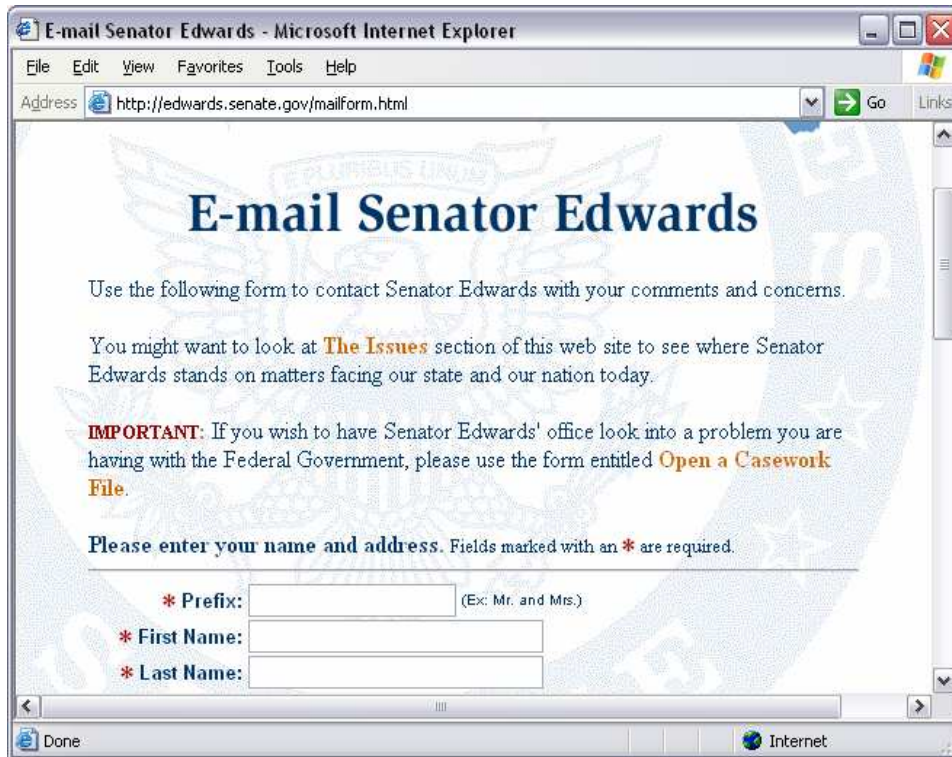


Figure 4 – Example of a vulnerable mail form on the Internet



Figure 4 – The post parameters for the mail form shown above.

In the example shown above the spam mail origin would seem to point to the machine that hosts the above mentioned script. Therefore the spammer will remain anonymous.

The third example of bulk mailing techniques is the transfer of mail via open proxies. In this case a virus (much like the Sobig virus) will be used to implant a mail sending program on an innocent victim's computer. The spammer will then remotely send spam via the compromised host. These machines are often referred to as "Zombie" hosts. Spammers use a technique called "Direct-to-MX" mailing, to send mail from a grid of these distributed "zombie" hosts. This technique will be discussed in the next section.

6 SPAM EVOLUTION AND TRICKERY

There are numerous techniques used by spammers to bypass filters and mislead mail users. These techniques are often very dynamic and evolve with changes in anti-spam technology. The paper will discuss a few of the most recent and prevalent tactics used by today's spam industry.

6.1 Misspelling common filter signatures

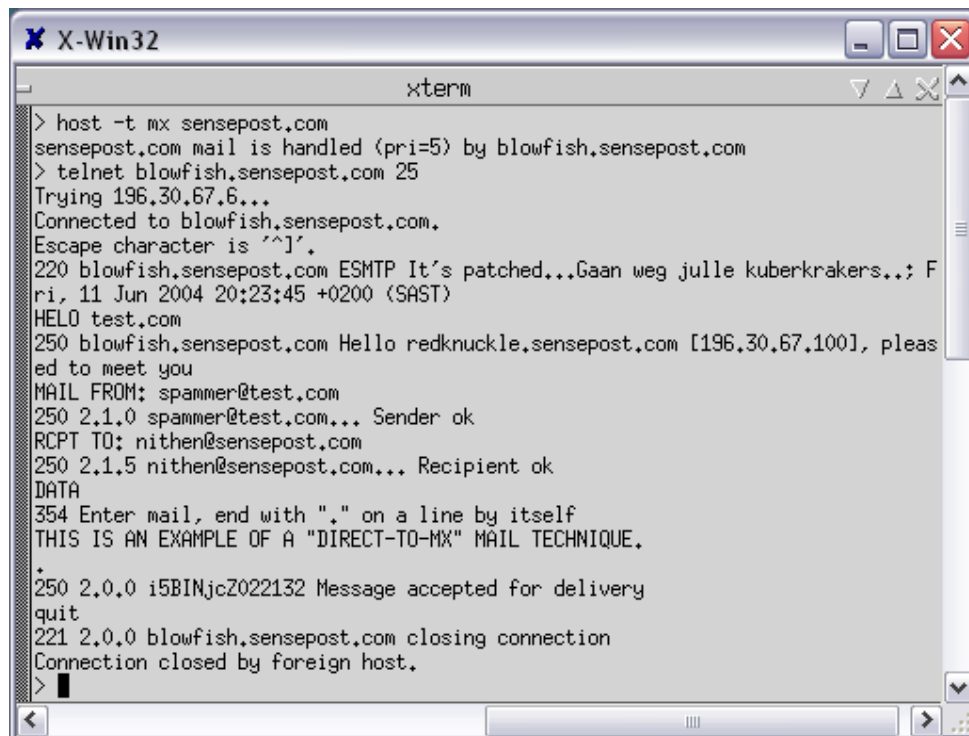
Spammers attempt to bypass simple filter rules by creatively misspelling or disguising certain words commonly used to tag spam by email filtering technology. (e.g. V1agr@)

6.2 HTML tag insertion

The email content bypasses spam filters by inserting HTML tags between known signatures (e.g. vigra) once the HTML is interpreted by the mail client the tags are either discarded or have no effect on the spammers intended content.

6.3 Direct-to-MX spamming

Spammers bypass the MTA process and connect directly to the MX of the domain using "Direct-to-MX" software. This aids offending spammers in hiding their activities from their ISP, forging mail headers and disguising the MX HELO.



```
X-Win32
xterm
> host -t mx sensepost.com
sensepost.com mail is handled (pri=5) by blowfish.sensepost.com
> telnet blowfish.sensepost.com 25
Trying 196.30.67.6...
Connected to blowfish.sensepost.com.
Escape character is '^]'.
220 blowfish.sensepost.com ESMTp It's patched...Gaan weg julle kuberkrakers..: Fri, 11 Jun 2004 20:23:45 +0200 (SAST)
HELO test.com
250 blowfish.sensepost.com Hello redknuckle.sensepost.com [196.30.67.100], pleased to meet you
MAIL FROM: spammer@test.com
250 2.1.0 spammer@test.com... Sender ok
RCPT TO: nithen@sensepost.com
250 2.1.5 nithen@sensepost.com... Recipient ok
DATA
354 Enter mail, end with "." on a line by itself
THIS IS AN EXAMPLE OF A "DIRECT-TO-MX" MAIL TECHNIQUE.
.
250 2.0.0 i5BINjcZ022132 Message accepted for delivery
quit
221 2.0.0 blowfish.sensepost.com closing connection
Connection closed by foreign host.
>
```

Figure 5 – The "Direct-to-MX" technique.

6.4 Invisible text

Bayesian filters weigh the presence of common spam signatures. This weight is measured against the entire mail content therefore the spammer hides non-spam related content in order to deceive the filter.

6.5 Random character strings (Hash-busters)

Many ISPs and large corporate companies use hash databases to identify known spam messages. Spammers insert random character strings into each message header to change the hash value of the message which would deceive such filters into believing the message is unique.

6.6 Embedding the recipient's email address with a hyperlink (Web-bugs)

This technique is commonly used by list makers. The embedded address helps them record mail responses. If there is indeed a response the embedded email is graded as a class "A" address implying the mail address exists and is in use.

7 PROACTIVE CONTROL MECHANISMS

The sections above suggest that Bayesian filters, ISP blacklisting, Hash databases and simple content filters are not going to prevent spam from getting to your mailbox. There is however proactive measures one can take to reduce spam. Using the information we now have, this section will discuss proactive techniques to reduce spam. The researcher believes that proactive controls integrated with reactive defences lead to a significant decrease in spam.

The best strategy for fighting spam would be to protect your email address from list makers. The paper has discussed common tools used by list makers to mine address. With this in mind the following controls were tested with great success.

7.1 Disguise email addresses available on the Internet

When posting or displaying your email address on the Internet, disguise it to ensure that it will not be mined by "spiders" or "robots" (e.g. example@domain.com – example at domain dot com) you need to HTML encode the address as well (e.g. exetc.).

7.2 Use pictures on your web site rather than text

On your corporate web site it is often necessary to provide clients with email contact addresses. Using pictures of the addresses rather than text would protect your contacts from automated spam tools.

7.3 Use secure mail-back scripts

Instead of giving clients a list of contact addresses use a mail-form. The script should be secure and the "mailto" addresses should not be posted as a hidden field.

7.4 Avoid short guessable addresses

Possibly reconsider address like charl@sensepost.com, and replace them with charlvdw@sensepost.com.

7.5 Bounce spam mail when possible

If mail is caught and tagged as spam and the return email address is valid (a mail sent by a list maker), send a generic "user does not exist" reply. This feature is often built in to many anti-spam and email technologies. Be careful not to bounce all spam messages this could result in a mail bounce storm or endless mail loop.

7.6 View all mail as plain text

This is a practice often suggested by many security professionals. With regards to spam, it will enable you to stop “Web-bugs” and other forms of HTML mining techniques.

7.7 Avoid giving out your email address to companies on the Internet

If given the choice opt not to give Internet companies your email address. Research has shown that many well known Internet company’s databases are bought by list makers. Setup a “spam” email account (nithen_no_spam@hotmail.com), and if you truly have no choice but to give an email address, use the spam address rather than your current email account.

7.8 Update filter and Antivirus signatures regularly

Updating the above mentioned signatures will help tag list maker emails and ensure that you are never the victim of a “mail via open proxy” virus.

8 CONCLUSION

Knowing the enemy is a prerequisite to successful victory. Spam costs businesses and the Internet a great deal of resources, and may in the near future threaten the very existence of email. The paper has acquainted the reader with spamming technology and methodology. This will empower the reader to initiate proactive spam controls to help fight the epidemic.

9 REFERENCES

APIG, Spam, www.apig.org.uk

Center for democracy and technology, Why am I getting all this spam, www.cdt.org

Dr Curtis Kret, Tracking Spam, www.securescience.net

Paul Graham, A plan for spam, www.paulgraham.com

PEW, Spam: How it is hurting email and degrading life on the Internet, www.pewinternet.org

www.mail-abuse.org

10 ACKNOWLEDGEMENTS

Thanks to the users of Darknet IRC channels and the USENET forums.

www.brightmail.com

www.spamassassin.org

www.spamcop.net

www.spamlaws.com

www.sampade.org

www.theregister.com