Typical spam characteristics

How to effectively block spam and junk mail

By Mike Spykerman – CEO Red Earth Software

This article discusses how spam messages can be distinguished from legitimate messages by looking at email headers and message content. It also mentions how spam can be blocked effectively by taking these typical spam characteristics into account.

Spam is not only offensive and annoying; it causes loss of productivity, decreases bandwidth and costs companies a lot of money. Therefore, every smart company that uses email must take measures in order to block spam from entering their email systems. Although it might not be possible to block out all spam, just blocking a large proportion of it will greatly reduce its harmful effects.

In order to effectively filter out spam and junk mail, we need to be able to distinguish spam from legitimate messages. To do this we need to identify typical spam characteristics & practices. Once these practices are known, suitable measures can be put into place to block these messages. Of course, spammers are continually improving their spam tactics, so it is important to keep up to date on new spam practices from time to time to ensure spam is still being blocked effectively.

Spam characteristics appear in two parts of a message; email headers and message content:

1. **Email headers**

Email headers show the route an email has taken in order to arrive at its destination. They also contain other information about the email, such as the sender and recipient, the message ID, date and time of transmission, subject and several other email characteristics. Most spammers try to hide their identity by forging email headers or by relaying mail to hide the real source of the message. Since they need to send mails to a large number of recipients, spammers use certain methods for mass mailing that can be classified as pure spam practices and can therefore be identified in the email headers. Although newsletters and legitimate mailings are also sent to a large number of recipients, these will generally not display the same characteristics since the message source does not need to be concealed.

Headers can also be used to trace back the origin of the spam message. However, in this article we are mainly focusing on how to distinguish a spam message from a legitimate message by looking at the email headers, rather than actually tracing the sender of the spam message.

Typical email header characteristics in spam messages:

- **Recipient's email address is not in the To: or Cc: fields:** The reason for this is that the recipient's email address is hidden in the Bcc: field or X-receiver field, along with a substantial number of other email addresses. Spammers do this in order to conceal the fact that the mail was sent to a large number of recipients, and presumably so as not to publish their email list. Some persons might add recipients to the Bcc: field for sending out 'legitimate' mailings, but these will tend to be of a more personal nature (which you might wish to block anyway) since most professional companies do not use this method for sending newsletters or mailings. Note however that if you do block emails without a local recipient in the To: or Cc: field, you will be blocking all bcc: messages.

- **Empty To: field:** This is also typical for spam messages. Because spammers send out bulk emails by entering all recipients in the Bcc: field or X-receiver header, the To: field is often empty. According to the RFC 822, Paragraph A.3.1. (http://www.w3.org/Protocols/rfc822/Overview.html), the worldwide standard for the format of email messages, every message is required to have at least one email address in the To: field. Therefore, if this field is empty, this must indicate 'shady practices'.
- **To: field contains invalid email address:** Instead of being empty or containing someone else's email address, the To: field can also contain a bogus email address, e.g. one without an @ sign or a non-existent one.
- **Missing To: field:** Emails that have no To: field at all, can quite definitely be considered as spam since this can only happen if done on purpose for spamming reasons.
- **From: field is the same as the To: field:** This is another common practice. Instead of entering a bogus or empty To: field, the email address in the From: field is also used in the To: field. Both email addresses are most probably fake email addresses.
- **Missing From: field:** Again the reasoning behind this is to disguise the actual sender of the message.
- **Missing or malformed Message ID:** Since the Message ID includes information about where the message is coming from, it is often missing or malformed (i.e. no @ sign or an empty string) in spam messages. The Message ID is in the form of xxx@domain.com. The first part can be anything and the second part is the name of the machine that assigned the ID. Although Message ID's are not strictly required, one can safely assume that they would only be missing or malformed if done deliberately to disguise the source of the message.
- **More than 10 recipients in To: and/or Cc: fields:** Many spam messages contain more than 10 recipients in the To: and/or Cc: fields. This can however also be used for 'legitimate mailings' but again these will tend to be of a personal nature (which you might wish to block anyway) since most professional companies do not use this method for sending newsletters or mailings.
- **Bcc: header exists:** In normal email messages, a Bcc: header does not exist since this is stripped from the mail.
- **X-mailer field contains name of popular spam ware:** The X-mailer field includes the name of the mailing software that was used to send the mail. If this header contains the name of popular spam software, such as Floodgate, Extractor, Fusion, Masse-mail, Quick Shot, NetMailer, Aristotle Mail, Emailer Platinum, Mast Mailer, The Bat and Calypso, this could indicate that it is a spam message. However, many spam mails do not contain an X-mailer header, or contain mail software that is widely used

such as Microsoft Outlook or Eudora. Since you might also be blocking legitimate mails if you do not filter on the right names, this header is probably not worth filtering on.

- **X-Distribution = bulk:** Spammers using Pegasus mail will have the X-header 'X-Distribution: bulk' added to their mail if it is addressed to a large number of recipients. This header occurs quite rarely, so you will not be able to catch large amounts of spam by filtering on this header.
- **X-UIDL header exists:** Incoming messages should not have an X-UIDL header since they are only intended for the mail server to stop it downloading messages more than once, for instance when 'leave messages on server' is checked. This header would normally be stripped when the message is received. Spammers add an X-UIDL header to try to get the recipient's mail server to download multiple copies of their message and therefore increase the chance that the message will be read.
- **Code and space sequence exists**: Many spam mails include a certain code for identification in the subject of the message. To hide the code from the recipient, a large number of spaces are usually placed before the code. This is done so that the recipient won't notice the code or that it is not displayed in the mail client before opening the message.
- **Illegal HTML exists**: Some spam messages include a code for identification in the text of the message. The text is entered outside the HTML tags so as to hide the code from the recipient. There is no reason to add text outside HTML tags, so the mere presence of illegal HTML can be treated as suspicious.
- **Comment tags to avoid detection by email filters**: Some spammers try to circumvent content filters by placing lots of HTML comment tags within the email body text. In this way, content filters will not recognize the spam words since they are separated by comment tags. The recipient however, will not see the comment tags since these are not displayed when viewing the message in HTML. Therefore it is important to use an email filter that can filter emails by removing HTML tags first.
- **HTML message without plain text body part**: HTML messages usually include a plain text version of the email so that recipients with email clients that cannot read HTML can still view the message in plain text. However, many spammers tend to send HTML messages without this plain text body part, not only to save on size but also to force recipients to read the HTML version. This enables spammers to embed links and unique IDs in the HTML code. For instance, many spammers include an image link that connects to a site when the message is opened. Since each message contains a unique ID, the spammer will know exactly which recipient has viewed the mail. In this way, spammers know how many people have viewed their message and which email addresses are still 'live'. When spammers know that your email address is 'live' this will entice them to send you even more spam, so it is important to put a stop to these kinds of spam messages by using a spam filter that is capable of checking this. Newsletters also tend to send messages without a plain text body part, so it is important to use a white list of allowed newsletters so as not to catch any false positives.

In a survey, Red Earth Software analyzed the headers of 500 spam messages and found that the following spam characteristics were mostly found:

Spam characteristics	**% of researched mails**
Recipient address not in To: or Cc: field	64%
To: field is missing	34%
To: field contains invalid email address	20%
No message ID	20%
Suspect message ID	20%
Cc: field contains more than 15 recipients	17%
From: is the same as the To: field	6%
Cc: field contains more between 5-15 recipients	3%
To: field contains more between 5-15 recipients	2%
To: field contains more than 15 recipients	1%
Bcc: field exists	0%
To: field is empty	0%
From: is blank or missing	0%

© 2002, Red Earth Software

2. **Message contents**

Apart from headers, spammers tend to use certain language in their emails that companies can use to distinguish spam messages from others. Typical words are free, limited offer, click here, act now, risk free, lose weight, earn money, get rich, and (over) use of exclamation marks and capitals in the text. Spam can be blocked by checking for words in the email body and subject, but it is important that you filter words accurately since otherwise you might be blocking legitimate mails as well.

How to stop spam

Now that we know the typical spam characteristics, how can we use these to stop spam?

Firstly, a mail filtering mechanism must be put in place to block out most of the spam and hoaxes coming into your organization. The email filtering system must be able to analyze email characteristics, classify a mail as spam, and either delete it, flag it (for instance add the word 'SPAM' to the subject), or quarantine it. Preferably, you will be able to make multiple filters that decrease in certainty whether a mail is spam. The more certain the filter is, the more drastic the action, for instance deletion of the message. If the filter can only indicate the possibility of a spam message, you could flag the mail or quarantine it. In order to avoid false positives, the email filtering system should be able to exclude white lists that for instance include allowed newsletters.

The email filtering system should filter out spam messages in three ways (in order of 'spam certainty'):

1. **Block spam at the gateway by checking domains in real time black hole lists:** There are a number of 'black hole lists' that contain IP addresses and domains from known spammers. By using these lists you can filter out a large amount of spam. Not only will you stop a large

proportion of spam messages from reaching your users, it will also save you utilizing your bandwidth to download spam messages since the message is blocked at the gateway, before the mail is even downloaded. There are two types of lists: (a) Lists of known spammer's domains, for example the Spamhaus Block List (SBL), and (b) Lists of mail servers that are open to relaying and therefore will allow spammers to send mail via their mail server. An example of this last kind of list is the Open Relay Database (ORDB). Whilst lists of the first type (spammer's domains) should be fairly accurate, lists of the second type, the open relay lists, can result in more false positives. This is because genuine persons that wish to contact your organization might not be aware that their mail server is being used for relaying. Therefore, it is important to treat each spam list differently. For instance, you could choose not to download all messages from domains listed on the Spamhaus Block List, and quarantine or delete (with the possibility to undelete) mails from the Open Relay Database.

2. **Filter out spam based on email header characteristics:** Most of the email header characteristics mentioned above can safely be used to classify a mail as spam. Therefore, you could decide to delete messages that contain any or some of the above mentioned spam headers. Since checking email headers is a fast process, it is good to check these before checking the actual email message content.
3. **Identify junk mail content:** There will still be spam messages that get through both filters mentioned above. The last way to distinguish these mails is by checking for spam message content. Depending on the words you select to filter on, this can usually be very accurate. For instance messages that contain phrases such as CLICK HERE, FREE!!, EARN MONEY, FAST CASH, BUY NOW, $$$, fast bucks and huge savings are almost 100% certain of being spam. Then there are words that could possibly be used in legitimate mails as well, such as money back, accept credit cards, credit profile, cash back, FREE. Therefore it is important to either perform different actions on the different sets of phrases, or to use textual analysis software that can minimize the chance of catching legitimate messages. For instance, by giving words or phrases a certain word score and specifying a word score threshold per email, you are able to specify quite precisely which messages should be blocked and therefore will decrease the amount of wrongly blocked messages. It is also important to apply case sensitivity to words, since spammers often use capitals in their messages.

Finally, you will need to educate your users. They must know that spam should be deleted straight away, and that they should never send a reply to a spam mail. This will just confirm that the email address is ‘live’ and will enable the spammers to sell the email address to other companies for further abuse. If the mail is a hoax, for instance a message about fake viruses, pyramid schemes promising lots of fast earned cash or victims asking for support by forwarding their mail, users should delete the message and not forward these mails. If users are educated in this way, you will be able to limit the negative impact of any spam or hoax message that has been able to pass your filters.

About the author

Mike Spykerman is CEO of Red Earth Software, a software development company that specializes in email policy enforcement software. The company's current products include Policy Patrol (www.policypatrol.com), an Exchange server and Lotus Notes add-on for blocking spam, viruses, offensive content, attachment quarantining, adding disclaimers and much more. Red Earth Software is a Microsoft Certified Partner.

References

RFC 822: http://www.faqs.org/rfcs/rfc822.html

E-Mail Spamming countermeasures: http://ciac.llnl.gov/ciac/bulletins/i-005c.shtml

Disclaimer

This article is in no way meant to provide spammers with tips on how to send out spam or bypass spam filters. Rather, it is meant to provide information for companies so that they can effectively block spam. By discussing spam characteristics openly, the author recognizes that spammers might be able to use this information in order to avoid email filters. However, in many cases spammers are already aware of the identifiable headers in their messages, whereas many companies trying to block spam are not. Therefore the author considers it useful to publish this kind of information.