



CMPE209 Research Paper

On

ANALYSIS OF EMAIL SPAM

October 14th, 2008

Under the guidance of Prof. Richard Sinn

Submitted By

Team Sparks

Gayatri Venkat

Hari Khanal

Tony Nguyen

Table of Contents

1. ABSTRACT.....	2
2. WHAT IS EMAIL SPAM.....	3
3. REASONS FOR SPAM.....	3
4. ANALYSIS.....	3
HOW SPAM WORKS.....	3
SOURCES OF EMAILS ADDRESSES.....	4
EFFECTS OF SPAM.....	5
ANTI SPAM MEASURES.....	5
5. HOW TO REDUCE AND AVOID SPAM.....	6
6. FUTURE OF SPAM.....	8
7. REFERENCE.....	9

1. Abstract

The paper details the current state of email spam. This form of unsolicited message is prevalent in Internet traffic and is a major source of concern to companies as well as normal users. This paper explains the many aspects of spams. Spam is a recent phenomenon started by the rise of the Internet and represents great potential financial gains for spammers. Spammers use various ways to gather email addresses and how their activities impact others. There is an ongoing competition between anti-spamming community and the spammers with many spamming techniques and anti-spamming methods. Finally, the paper offers some advice on how user can protect themselves from spam and reduce email pollution.

2. What is Email Spam

Spam is misusing the internet with the numerous copies of the same message. It is an attempt of forcing users to receive messages that are not requested by them. Email Spam involves sending direct emails to users emails address. Spammers retrieve the emails addresses from various sources and send numerous unwanted mails in bulk to all the email addresses at a time. These spam mails are called Unsolicited Bulk Email (UBE) or Unsolicited Commercial Email (UCE), that usually contain commercial content w.r.t advertising, product schemes, etc.

Email Spam started in 1990 when Internet was made available to the general public. It has grown to such an extent that spam emails comprise up to 80 - 85% of the emails in the world.[1] It has been successful in confusing, frustrating and annoying millions of users till date. It is difficult to define spam but it follows "You know it when you see it" phenomena. Since then spammers have been invading users' inbox with bulk messages containing amazing product deals, adult websites and "get rich overnight" schemes and offerings and letters which have not been subscribed by the user by choice.

3. Reasons for Spam

The main motivation for spammers is money and the cost involved in sending mails. Spammers send millions of mails in the hope that few people will respond. Even if the response rate is low, the spammers still benefit as the cost of sending those millions mails is still zero. For example, if a person wants to sell a product for 1\$ and he sends a mail with this offer to 1 million people. Even if 0.1% of the people response to the offer, the spammer earns \$10000 without incurring any cost. All that a person requires is a list of valid email address, spamming software and email server and the process is cheap, fast and simple.

In addition, there is a lot of cost involved in the other means of advertisements like print ads, newspaper, etc. This makes advertisers discriminate among the recipients as well as the volume of ads sent out as each additional ad incurs an additional cost. In case of email spam, the cost involved in sending the first email is the same as the one millionth email.

4. Analysis

4.1 How Spam Works

There are three parties responsible for a spam email to reach one's inbox. They are as follows: [2]

1. Advertisers: The main cause of spam is the need for someone to advertise about a product that he intends to sell. Some of them are spammers themselves who know how to send spam mails to everyone. While the remaining are just ordinary people who are computer illiterate who hire some third party

experts to send spam mails on their behalf. These advertisers are responsible for the content of the mails and the point of contact for user response.

2. Spam Service Providers: The Spam Providers are the people who possess the required knowledge, expertise, hardware and software to send out a million emails. These providers charge the advertisers for sending spam emails on behalf of them. Even if the response rate for the advertisement is 0%, the spam Service Provider earns his share of profit.

3. Spam Support Services: The Spam Support Services include the web site hosting companies and ISPs located in countries where spamming is legal. Spam Support Services find ISPs willing to offer support services in countries like China, Argentina, Brazil.

4.2 Sources of emails Addresses

The most common question that worries users when they receive a spam mail is how do spammers get their email address when users have not given it by choice? Spammers get valid email addresses from a number of sources, an incomplete but most typical sources include:

1. **Newsgroups and Chat rooms:** People many a times use their screen names or specify their email addresses in newsgroups and chat rooms. The spammers easily get the email addresses by using their software.
2. **Websites:** Spammers can easily run a program that would fetch email addresses from websites containing any reference email addresses. In addition spammers can also fetch addresses of users participating in forums and discussions on any website.
3. **Email Servers:** Spammers can obtain email addresses from large email hosting companies like Yahoo, Hotmail, MSN, etc. They use a dictionary attack that uses a software to open up to the target mail servers and sends multiple random email addresses. The “live” addresses are then recorded by the spammer's software. This list is then shared with other spammers.
4. **Advertisement:** The most commonly used method is sending attractive product schemes to users through emails. These mails often tempt users to specify their email addresses and thus spammers get valid email addresses of such users. In addition to this, users sometimes unsubscribe to offers or newsletters which make their email addresses available to spammers.
5. **Email addresses from Companies:** Email addresses of numerous users are sold to spammers by many companies in the form of CDs.
6. **Compromised computers:** Hackers can supply addresses (address books or documents with email addresses) from compromised machines to spammers.
7. **Social engineering:** The spammer uses a hoax to convince people to supply email addresses. An example is an email detailing a sick child whose life depends on donation from an organization, cc'ed on the mail, which will donate money based on the number of people who get the email. Since the organization is the spammer, users are willingly giving up email addresses every time the mail is forwarded.

4.3 Effects of spam

While most of us receive several spams a day, the effects of spams are much more serious when looking at large scale.

1. **Very costly:** Spamming is the only form of advertisement where the cost almost entirely incurs on the receivers. Spammers can generate millions of emails a day from a single computer with minimal cost, Internet connection fee. Multiply this single spamming computer by thousands or millions and the amount of spams on the Internet is unimaginable. The receiving mail servers have to spend CPU cycles filtering all incoming mails for spam. Filtering a spam sent to several thousand accounts may take several seconds on the server. The spam machine which sends millions of spams a day would require several mail servers dedicated to just filtering spams. Storage is another very expensive aspect of spams which in many cases contain images. If spam filtering isn't enforced, it's conceivable that terabytes of wasted storage on spams. Thus, spam for companies is a major source of wasted resources in terms reduced service quality and availability.
2. **Unusable mails:** Ineffective anti-spam software or unfiltered mail accounts may well be unusable if there's a high ratio of spams to real mail. As a experiment, a Hotmail account with no filtering would get an average of thirty or more spams. This number can be significantly higher if the mail server didn't have anti-spam software. A mail user who sees one real mail out of thirty or fifty messages would be very weary of this modern communication technology.
3. **Unsolicited worthless information:** This is the most obvious negative aspect of spam as the contents are typically useless to 99.9% of the receivers. To most users, dealing with spam is a time consuming task that yields no benefit except frustration.
4. **Hacking:** This illegal practice is used to gather email addresses or to send deceive mail servers by sending spams from compromised computers.
5. **Illegal spam:** While most spams are quite harmless in the sense that they are marketing attempts. Some spam contents are considered illegal and may cause trouble to innocent receivers. Examples would be porn solicitations to users in countries where pornography is illegal and/or solicitation to political sites that are considered illegal by the receiver's ruling government.

4.4 Anti-spam measures

The ongoing battle between anti-spam community and spammers is similar to one between cryptanalysis and cryptography. Below are details of anti-spamming techniques to block existing spams and how spammers get around these techniques.

1. **Keyword detection:** Email service provider or mail client used certain text phrases to identify and block incoming mails. This mechanism is effective but can also produce false negatives where legitimate mails also contains blocked phrases. Moreover, spammers can easily spoof this detection by an altered phrases (e.g. 'viagra' to 'v1agra') or simply sending spam message embedded in an image rather than plain text.
2. **Blacklist** - Email servers keep a list of IPs from which mails are automatically denied. This is an efficient way to block spams from known spamming sources without spending CPU cycles to examine the contents. The drawback is constant updates of this list to include new spammers. However, spammers can constantly update their ISP accounts to change IPs address which

renders the blacklist ineffective. Additionally, spammers can use a mechanism called Botnets where a network of compromised computer can be controlled for certain activities, in this case sending spams on behalf of the spammer. This is a cooperation between hackers and spammers to compromise machines from which they send spams. Blacklist are either ineffective against this technique or marking these IPs as false negatives (spammers) and block all mails from these clients.

3. **Keyword detection using OCR:** Anti-spam software performs keyword detection with text derived from images. However, spammer can fuzzy (include noise and distortion) the images to make it very computing intensive, impractical, to get text from the images.
4. **No mail policy:** This is a possible final anti-spamming techniques that eliminates the use of emails thus accepting no emails. In these cases, organization replaces email communication with online forms where a sender will explicitly fill out and submit to transmit a message to the intended receiver. While this is a strong solution, it suffers from usability concerns.

Furthermore, there are new spamming trends that represent huge challenges to the anti-spamming community.

- **Spam as attachment:** This approach keys on the fact that most people expect spams to be text and attachment of certain types are typical important, business document. Anti-spam software are typically written to analyze the message content but not the attachments. With this information, spammers now send mails with the message attached as either a PDF or Microsoft Excel document. Since businesses consider PDF and Excel files as the common method of exchanging official documents, users typically need to open the attachment, get spammed, before they can determine whether or not the attachment is a spam. At the moment, only some anti-spam software can filter PDF and Excel contents.
- **Package in zip file:** Enclosing a spam in a zip file is one of the new trends to make it harder to inspect attachment. Anti-spamming software would need yet another step to get the content before filtering.

5. How to Reduce and avoid SPAM

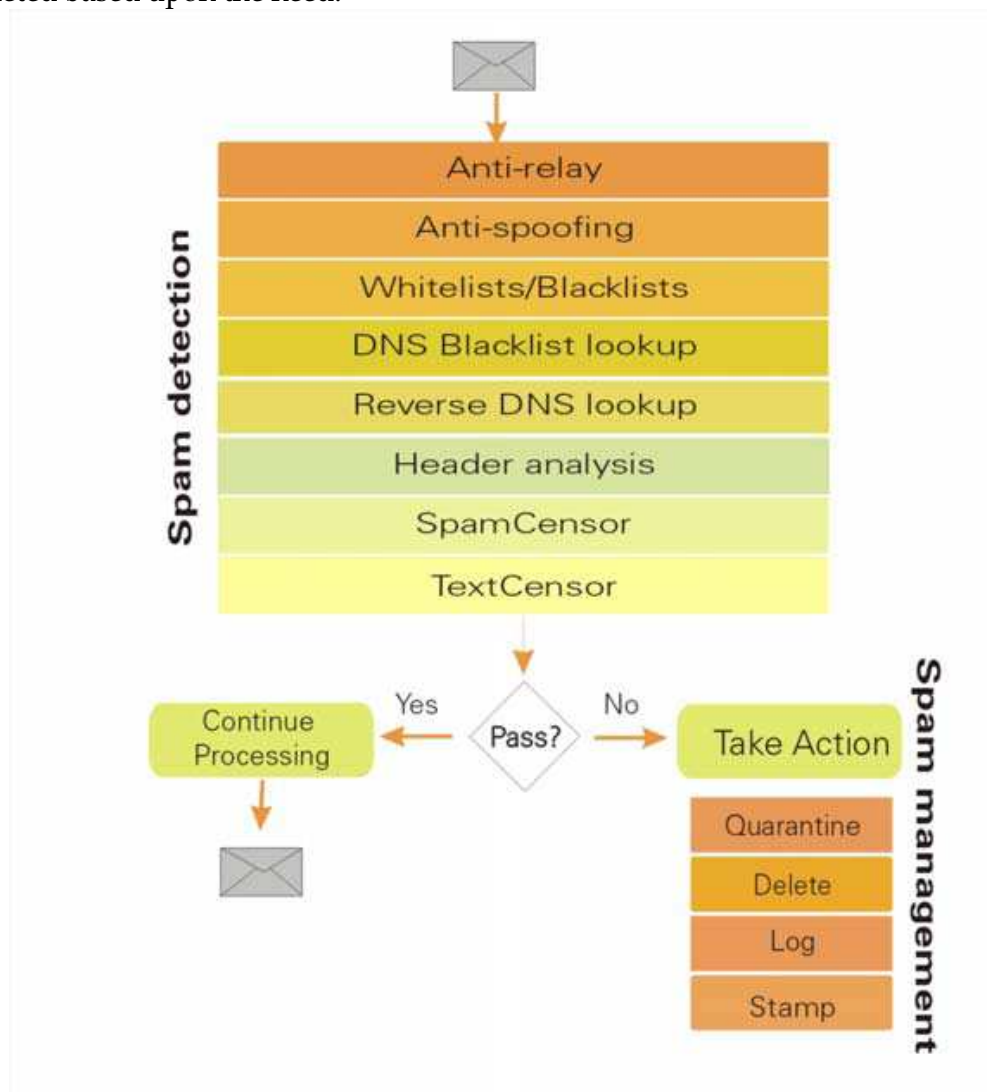
There are several ways of approaching anti-SPAM; however, a combination of them is better than a specific anti-SPAM approach. In context of email-SPAM, an infrastructure can be designed to detect and remove SPAM before emails get to mail server. For example: MailMarshal Gateway server that can remove SPAM before emails reach mail server. [4]

The architecture of MailMarshal is simple. Commonly, there is firewall on the frontend of the corporate network; however, firewall is an optional system and is not required for MailMarshal. Once the traffic including email traffic passes through firewall, email traffic enters the MailMarshal so that SPAM can be detected and managed. MailMarshal is simply a content filtering gateway for SMTP. Mail traffic defined as spams are filtered and other allowed mail traffic finally gets to mail server. All the clients connect to mail server to receive email.

MailMarshal is multilayer system for detecting and managing SPAM. A few well known features are advanced text analysis, message header analysis, host Analysis, domain or user blacklists, domain or user whitelists, anti-relay, anti-spoofing of local email, DNS blacklists, MAPS RBL, reverse DNS lookups, reporting. [4] Some of the features are described below:

1. Advanced text analysis allows scanning the text in the email. Based on a characters or array of characters, the email can be considered SPAM. For example: if text include "Free \$\$\$," it can be categorized as an SPAM email and can be blocked before reaching email server.
2. Message header analysis allows scanning of the header similar to message text. Based on the known texts categorized as SPAM, the email can be allowed or trashed. Mail server does not get the email.
3. Host analysis is based on hostname or equivalent. Suppose SPAM emails are coming from a specific host or known hosts, the email can be categorized as SPAM and trashed before reaching email server.
4. Domain or user blacklists allows making decision based on domain name. A domain can have multiple hosts, so it becomes efficient in blocking SPAM email from hosts from a certain domain name. It also reduces time and configuration information using domain name in addition to host name.
5. Whitelists are based on exception in terms of domain name or hostname or user. In order to exclude certain hosts and users from a domain, whitelists are very useful and provides solution.

Since MailMarshal is multilayer approach, email traffic goes through each layer and based on each layer policy, email are trashed or allowed. After mail traffic passes through first layer of spam filtering, it moves to next layer. New layers can be added and inserted whereas existing layers can be modified or deleted based upon the need.



MailMarshal Operation

MailMarshal is a very good approach because it is transparent to client. Clients do not know about the Mail marshal. So, users without strong knowledge can be accommodated in this approach. Besides, the policy applied on the MailMarshal is uniform for all the users. It doesn't require specific set of policy for each user and multiple copies of the policies. Furthermore, it provides scalable approach and policies can be added or deleted or modified as needed.

The best scenario is to have server-side spam control like MailMarshal and client-side spam control. On top of the corporate solutions like MailMarshal, email software like MS Outlook and online email like Gmail also have anti-SPAM features built-in. This allows the end users to have additional control on spam. The common way to get client software up-to-date is by updating and upgrading them as updates becomes available.

MS Outlook is client-side software to manage and store emails locally. Depending on mail protocol used, client policies can be synchronized with server. For example: Microsoft Exchange Server is a common mail solution. The MS outlook can store emails locally and synchronized with MS Exchange Server. Protocols such as IMAP can also synchronize client and server so that client and server have same policies and settings. Several options are available to control spam. Filters can be applied based on trust (safe list) or junk policies. [8] The dark side when applying these filters is that some of the mails that are not junks also are marked as junk and sent to Junk folder.

On the other hand, Gmail uses authentication to validate email and uses OCR. [9] Gmail claims to have decreased the number of spam making to inbox folder through rigorous check on all senders through authentication. [9]

So, a combination of server-side along with client-side spam filtering is the best solution.

6. Future of SPAM

Spammers have been ahead of anti-SPAM solutions. The reason is anti-SPAM solution is based on existing SPAM. It has not been possible to provide anti-SPAM solution that can work on future SPAM. Looking at the history of SPAM, it can be predicted that SPAM will keep on existing. However, SPAM can be limited by providing solution for existing spam and detecting and blocking new spam from time to time.

Spammers use multiple hosts on the network clouds to send spam more efficiently. The intention could be either business to steal information. The node in internet that process and distribute SPAM is also zombie network. At the beginning the spammers used test in message and header. The solution that can analyze the text header and message were deployed. Then, spammer used image instead of text and the existed solution did not work. A new solution of OCR (optical character recognition) was provided to turn the image to text. [5][9]

Spammers have kept on changing format of the attached file to the email so that existing system become unable to detect SPAM. From PDF to ZIP to different format, the attached file format keeps on changing. [5]

It is impossible to block all the SPAM in future because the new category of SPAM keeps appearing in the internet. The only efficient way to block SPAM email efficiently is to use scalable solutions like

MailMarshal. Deploy and modify the policies as needed to fit the need of the organization.

On the other hand, the scalable solution is going to cost more compared to client-side based solution. The presences of dedicated anti-SPAM servers can help to eliminate known SPAM from getting to mail server. However, normal users and small companies may not be able to deploy dedicated servers for anti-SPAM solutions. But ISP can control for normal users but it might not be efficient compared to organization dedicated spam filtering system.

For enterprise level, it becomes feasible to deploy dedicate anti-SPAM server solution. The servers policies can be modified and added as needed. In addition, the SPAM does not require hardware cost overhead. So, spammers have been able to increase number of SPAM exponentially. Through different software and scripting language, spammer can easily send spam to multiple users at once or send spam to each user one at a time. Spammer can also send spam with fake hostname or domain to end user.

It has become really challenging to provide solutions. Moreover, the overhead cost to deploy the expensive anti-spam solution may not be affordable for mid-size organization or regular user. Furthermore, email distribution list makes it easier for a spammer to send a single spam to multiple users. A distribution list can include multiple email addresses that can be distribution lists and end user email addresses. Most likely, every organization has a mail distribution alias that includes all the employees. If spammer gets this mail distribution alias, then a spammer can flood the organizational network. This is another issue that will keep on existing in future.

In today's context, it is likely that spam is going to exist in the future. As the number of internet users is growing, so the number of spam. As developing nations like China and India are growing rapidly in internet, spammers and anti-spam solutions are going to increase too. This year, China has surpassed the USA in the number of internet users.

7. Reference:

- [1] http://www.maawg.org/about/MAAWG20072Q_Metrics_Report.pdf
- [2] http://media.wiley.com/product_data/excerpt/56/07645596/0764559656.pdf
- [3] http://wnd.com/news/article.asp?ARTICLE_ID=32928
- [4] http://download.netiq.com/CMS/WHITEPAPER/NetIQ_Controlling%20SPAM%20White%20Paper.pdf
- [5] <http://www.gfi.com/whitepapers/attachment-SPAM.pdf>
- [6] <http://www.microsoft.com/protect/yourself/email/default.mspix>
- [7] <http://computer.howstuffworks.com/SPAM1.htm>
- [8] <http://office.microsoft.com/en-us/outlook/HA011590551033.aspx>
- [9] <http://googlesystem.blogspot.com/2007/10/how-gmail-blocks-spam.html>