Author(s): Malik, Gaurav
Article title: Tackling Spam and Spoof Email
Year of publication: 2006
Citation: Malik, G. (2006) ‘Tackling Spam and Spoof Email’ Proceedings of the AC&T, pp.65-71.
Link to published version:
http://www.uel.ac.uk/act/proceedings/documents/ACT06Proceeding.pdf

TACKLING SPAM & SPOOF EMAIL

Gaurav Malik
Innovative Informatics Research Group
gaurav@uel.ac.uk

Abstract: The loss of productivity due to Spam has reached a critical limit. Spoof emails have dented confidence of people in communications from organisations. This is happening in an age where email has been recognised as a cost effective way of communicating. Companies have to invest resources to increase the confidence of consumers rather than abandoning the use of emails. This leaves two avenues of pursuing the matter, either email vendors have to implement safeguards or users have to implement technology and procedures. The paper will look at ways in which spam and spoof emails are being tackled and also make suggestions on how confidence can be raised by the use of hybrid approaches.

1. Introduction

Email is now a killer application used by companies for communication and marketing. As a non secure delivery method it is susceptible to abuse. Many solutions have been proposed from content of the email and the domain of the sender. None of the systems seem to be fool proof and have their own limitations. The decision now comes down to the sender and receiver on what technology to implement.
No system is fool proof, it is also interesting to note that the top 5 email domains are aol.com, hotmail.com, yahoo.*, earthlink.com & gmail.com all implement a combination to spam & spoofing technology.

2. Cost of spam

Spam and spoof email have reached a stage where it is an epidemic. It is an unusual day when you open an inbox and there is no spam email in your mailbox (Cerf, 2005 p39-43). The cost of spam can be worked out through a simple calculation (CMS Praetor):

Workplace and email environment

- Number of Employees with email: 100 (a)
- Number of Workdays per year per employee: 230 (b)
- Average hourly salary per employee: £10 (c)

Assumptions about email usage

- Average number of spam emails per day, per employee: 25 (d)
- Number of seconds wasted with each spam message: 5 (e)

Total corporate cost of spam:
Lost Salary

- Yearly: £7986.11 (f) (a*b*c*e)/3600
- Daily: £34.72 (f/230)

Lost Productivity:

- Yearly: 52.81 Days

Cost of spam for each employee
Lost Salary

- Yearly: £79.86 (f/a)
- Daily: £0.35 (f/a*b)

Lost productivity:
Yearly: 12.67 hours per Employee

The figures here represent a typical SME. If we increase the figures, we will see a linear increase in the costs and hours involved. The figures are frightening. On the other hand direct email marketing has replaced direct mail as a vehicle to attract customers. In research published by Gartner (GartnerG2, 2002), it states that email marketing is set to overtake direct mail. Use of email as a marketing tool, is beyond the scope of this paper (10 rules for successful permission-based e-mail marketing, 2005) It would be useful to point out techniques and development in the world of email marketing.

Use of a clean prospect database.

Due to attrition and change in people's email addresses a prospect database can get out of date. It is considered prudent to clean the database to ensure the number of bounce backs decrease. In many companies invalid emails are sent to a "postmaster" account. The receipt of many emails can result in a sender's address being black listed.

Opt-in methodology\ Permission based marketing.

Most sites are now required (in some countries by law) to present users with an option to opt-in. This enables the company to only have users who have explicitly chosen the option to receive information.

Emails must also present subscribers the ability to unsubscribe via a simple method.

Use of email as an effective and convenient communication cannot be ignored. Now it remains a matter of identifying the email as genuine.

3. Current approaches in identifying spam.

3.1. Content Based Analysis

Many papers have been published where novel approaches are being applied to identify spam. These include the use of neural networks (Chuan, Xianliang, Mengshu & Xu 2005 pp 34-39), Cluster Analysis (Deshpande, Deshpande, Bhuleskar 2005 pp 103-109). Essentially it all boils down to analysis of the email received.

The problem with the above approaches is that the email has been received by the recipient's mail server before action can be taken. This creates unnecessary load on the server. The above approaches are novel but need to be implemented in mail servers. At present mail servers identify spam through the following methods:

- Keywords matching
  This method looks at the subject of the email and the body of the email. Majority of spam use the same words. One of the problems with this method is that it requires manual updating. Also legitimate email can get labelled as spam due to part of a string being identified. For example "sex" is a substring of Middlesex.
- Bayesian analysis
  Bayesian filtering is based on the principle that most events are dependent and that the probability of an event occurring in the future can be inferred from the previous occurrences of that event. (GFI)
  - It examines the entire message and looks at words in their context to the message as opposed to keyword matching.
  - It is constantly learning and self-adapting. It also learns from new spam and from valid incoming and outgoing emails. The Bayesian filter is also designed to adapt to new spam techniques. It is also difficult to fool.

    - It is sensitive to the user. By reading valid emails it learns about the user profile and then can better judge whether an email is spam or not..
    - It is multi-lingual and international. Keywords are available in various languages.
- Whitelist/Blacklist
    - Whitelist are hosts from who we wish to receive emails. This can be done manually. There are some email servers which can automatically add outgoing hosts to their database. Blacklists are hosts from whom we do not wish to receive email.
- Mail header analysis
    - Mail header checking looks for anomalies in the email header. Missing "from" addresses and multiple recipients etc.

3.2. Limitations of content based approaches

Limitations of the above methods stem from the fact that corporate mail filters cannot risk labelling valid emails as spam. So they take a cautious approach when labelling email spam. All the above methods require the system to learn. Initially the number of spam email is quite high, but it will gradually grow small as the rules are learnt by the mail server.

Current approaches in removing spam lies with the recipient mail server. The outgoing server and all other relaying servers do not play any part in identifying the email and spam. RFC 821, RFC 822 requires the presence of a "POSTMASTER" account (IETF 1997). This account is meant to be used as a catch-all for all emails with invalid email addresses for a particular domain. This is an additional load which has to be performed. Common mistakes may be "gaurav@uel.ac.uk" being spelt as "guarav@uel.ac.uk". The provision of the postmaster account is to ensure that the email is delivered to the correct address.

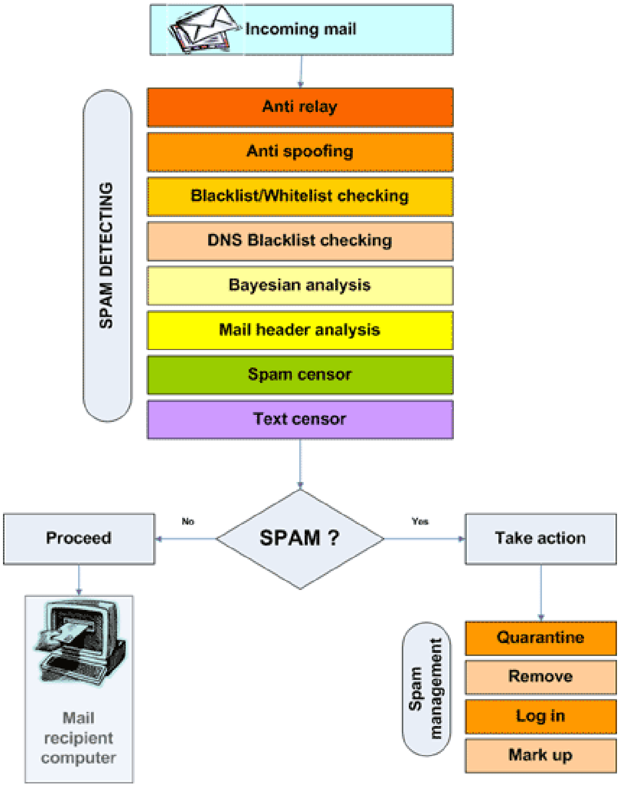

Figure 3 Common methods in which conventional emails servers work (MSExchange.org 2004)

3.3. Domain Based Analysis

SPF (Sender Policy Framework)

SPF works more on the principle of email authentication rather than spam prevention. Companies use SPF to publish the addresses of their email servers. The recipient server can then check these records to ensure the mail is coming from the addresses which are published.

SPF is not automatically designed to prevent spam rather it proposes the creation of a reputation service (Wong 2004).

Most DNS services support the publication of SPF records. Various mail server programs support SPF checking. Some of these are the most popular Mail Transfer Agents (MTA's) SendMail, Postfix, Exim, Q-Mail, Courier, Microsoft Exchange, Sandtronics WildCat (SPF).

However, recent reports show that 34 percent of SPF-registered mails were spam (Neil & Veitch 2004). So SPF on its own is not deemed to be the silver bullet. On the other hand in June 2005, the IETF accepted the SPF specification for RFC status (Wong & Schlitt 2005).

Yahoo Domain Keys

Yahoo DomainKeys (Yahoo Domain Keys) uses an RSA public/private key method. All outgoing emails are signed with the private key. The public key of the domain is stored in the DNS, where it can be checked. It requires both receiving and sending email servers to implement the technology. DomainKeys work in a similar manner to SPF in that it deals with email authentication.

Domain Keys has advantages over SPF where email which have forwarded by external relays and forwarding services can be verified with Domain Keys. Domain keys also ensure the integrity of the email contents which SPF doe s not.

Cisco and Yahoo have partnered together to create a new standard called DomainKeys Identified Mail (DKIM) which has presented to the IETF for consideration as a new e-Mail standard to address E-Mail forgery and phishing in July 2005 (DKIM).

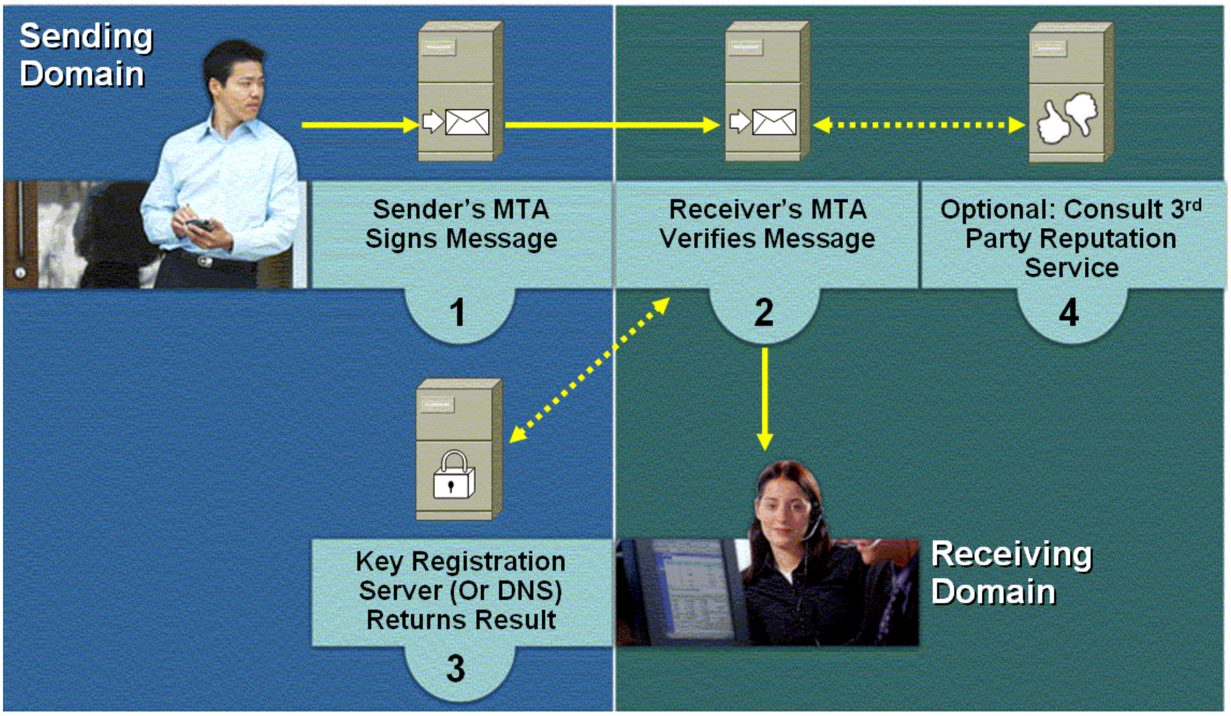
Sending Domain
Sender's MTA Signs Message
1
Receiver's MTA Verifies Message
2
Optional: Consult 3rd Party Reputation Service
4
Key Registration Server (Or DNS) Returns Result
3
Receiving Domain

Figure 4 DomainKeys Identified Mail

Microsoft Sender ID

Microsoft initially labelled its technology Caller ID (Sender ID 2005). In May 2004, Microsoft submitted its proposal along with SPF to the MTA Authorization Records In DNS (MARID) working group to develop an Internet standard. Sender ID validates the actual "From" address of an email, unlike SPF which only validates the originating email address of the message.

4. Conclusions

DomainKeys Identified Mail (DKIM), Microsoft Sender ID & SPF are the prevalent technologies for identifying spam mail. All three have their pro & cons, yet it seems that rather than agreeing a common working ground all three technologies have gone their separate ways. Yahoo has implemented DKIM in its mail service and also provided the technology for Google's GMail. Microsoft uses Sender ID for its own Hotmail service and has built it into its own Exchange Mail server. It seems ironic that in June 2004 when these companies met as part of Anti-Spam Technical Alliance (ASTA 2004), the recommendations they made were more on best practice rather than actual technology implementation.

Microsoft to its credit did try and take its Caller ID technology forward with SPF and present it as Sender ID. Unfortunately due to patenting issues the alliance was broken. Both Microsoft and Yahoo due to their own popular web based email service are the victims of spam, have tried to implement their own proprietary email system. While trying to develop protection for spoofing and spamming they have invented and in some cases innovated existing ideas. As there is currently no RFC for spam they have tried to push their ideas to be adapted as a standard. After the initial dabbling with the IETF, where neither companies succeeded they now seem to have taken a rest. SPF an open standard is now every close to being made into an RFC standard.

Domain based email analysis and content based email analysis are the two tools which can be used to fight spoof and spam emails. Most of the above technologies add work to the recipient server (DKIM, adds load on both servers). The recipient server can shed some of this load through the caching of records.

Spoofing can be tackled only at the recipient's end. We can check for sender authentication once the message has been received.

All spam mails generate from users who have an ISP account, whether it be broadband or through hosting. It would be encouraging to see some work being done where the email is rejected from the sender's SMTP server. This is the only time where the message is seen in its entirety before it starts its packet journey using TCP\IP. We can then identify users who are using their accounts for spam.

It seems now that a reputation and accreditation system would be the best way to tackle spam. Authentication only tells us who the sender is. Reputation systems maintain safe lists which are then used by the receiving companies. Companies like Microsoft are now using reputation systems like Habeas, Bonded Sender & GoodMail to be able to better identify email. Reputation & accreditation helps us present "false positives" where legitimate email gets identified spam. This however makes more sense for organisation which sends mass mail like email marketers, newsletters. Also the cost of sending emails is now being passed to the sender. The sender has to now participate with one of the reputation services, to ensure that emails reach their recipient. The problem faces smaller companies who wish to use email as a marketing tool.

It would seem now like all authentication systems on the Internet we have now created a new business model for Reputation & Accreditation services. The move seems to be towards a paid email service where companies which need to send mail for marketing will subscribe to a reputation service. Effectively spam will be priced out due to the costs involved.

5. References

10 rules for successful permission-based e-mail marketing (2005) [online], Available from: http://www.microsoft.com/smallbusiness/resources/marketing/online_marketing/10_rules_for_successful_permission_based_email_marketing.mspx [Accessed 23 Nov 2005]

Anti-Spam Technical Alliance (2004), Available from: http://www.microsoft.com/presspass/press/2004/jun04/06-22ASTAPR.mspx, [Accessed 24 Nov 2005]

Ashutosh Deshpande, Adwait Deshpande, Richard Bhuleskar; Cluster analysis for spam E-Mail filtering, Proceedings of ICGES-05, 103 – 109.

CMS Praetor, cost of spam calculator , http://www.cmsconnect.com/Marketing/spamcalc.htm,

David Neal & Martin Veitch (2004), IT Week 05 Sep 2004, [online]; Available from: http://www.itweek.co.uk/itweek/news/2085609/sender-id-gains-support; [Accessed 20 Nov 2005]

DomainKeys Identified Mail (DKIM), (No Date), [online], Available from: http://mipassoc.org/dkim/index.html, [Accessed 24 Nov 2005]

GartnerG2 (2002) Says E-Mail Marketing Campaigns Threaten Traditional Direct Mail Promotions, http://www.gartner.com/5_about/press_releases/2002_03/pr20020319b.jsp

GFI (No Date) White Paper [online]; Why Bayesian filtering is the most effective anti-spam technology

IETF (1997), Mailbox names for common services, roles and functions; Available from: http://www.ietf.org/rfc/rfc2142.txt [Accessed 23 Nov 2005]

M.Wong, W.Schlitt (2005); Sender Policy Framework (SPF) for Authorizing Use of Domains in E-MAIL; Available from: ftp://ftp.isi.edu/internet-drafts/draft-schlitt-spf-classic-02.txt

Meng Weng Wong (2004), The Aspen framework on Reputation and Accreditation; [online]; Available from: http://archives.listbox.com/spf-discuss@v2.listbox.com/200406/1268.html [Accessed 20 Nov 2005]

MSExchange.org (2004), Four Popular Anti Spam Filters for Exchange Reviewed [online]; Available from: http://www.msexchange.org/tutorials/Preventing_Spam_Antispam_Filters_MS_Exchange.html [Accessed 21 Nov 2005]

Sender ID (2005), [online], Available from: http://www.microsoft.com/mscorp/safety/technologies/senderid/resources.mspx, [Accessed 24 Nov 2005]

SPF project (No Date), Available from:http://www.openspf.org/; [Accessed 20 Nov 2005]

Vinton G. Cerf (2005), Spam, Spim and Spit; Communications Of The Acm April 2005/Vol.48,No.4, 39-43

Yahoo DomainKeys (No Date), Proving and Protecting Email Sender Identity; [online], Available from: http://antispam.yahoo.com/domainkeys, [Accessed 24 Nov 2005]

Zhan Chuan, Lu Xianliang, Hou Mengshu, Zhou Xu (2005); A LVQ-based neural network anti-spam email approach, ACM SIGOPS Operating Systems Review archive Volume 39 , Issue 1 (January 2005),34 - 39