Available online at http://acfa.apeejay.edu

APEEJAY JOURNAL OF COMPUTER SCIENCE AND APPLICATIONS

ISSN:0974-5742(P)
Vol. 2, January 2014

METHODS TO PREVENT HARVESTERS BY REMOVING EMAIL ADDRESSES FROM SPAMMERS LISTS

Rajiv Mahajan[1] Gagandeep Singh[2]
[1]*Global Insitutes of Managment and Emerging Technologies, Amritsar*
[2]*Sai Technology Campus, Amritsar*

Abstract

To prevent email from being crowded out by spam, there are several concepts: Some approaches try to remedy symptoms and filter spam out of the inbox. Some try to fight spammers using tar pits or sue them. Others describe solutions like obfuscating email addresses, so spammers would not know whom to send their spam. This paper discusses that there are methods to have one's email address removed from spammer's list and Thus reduce the overburden over the network and hence improve the Bandwidth of Existing network Improvement in the It is based on an ongoing real world experiment and their

Keyword: Security, Preventing harvester, Obfuscation, Tarpit, Iperf, SMTP

1. INTRODUCTION

Spam presents a significant challenge to users, Internet service providers, states, and legal systems worldwide. The costs of spam are significant and growing, and message volume threatens to destroy the utility of electronic mail communication. Postini, Inc., a provider of e-mail security services for corporations, reported that only 14% of the seven billion messages it processed in November 2009 was legitimate e-mail. Estimates of the worldwide cost of spam to consumers and businesses reach as high as $20 billion each year. These costs include: time spent deleting unwanted messages; the cost of software, hardware and services designed to block unwanted spam; the cost of bandwidth to handle transmission; and the costs for server storage and processing required to deal with the unwanted messages. In addition, there are intangible costs: legitimate messages that are erroneously filtered by spam blocking systems; loss of efficiency as users turn to less useful communications mechanisms when e-mail becomes unreliable; and the harm to Internet domain owners' reputations whose domains are spoofed in fraudulent e- mail. Finally, a significant percentage of spam promotes some type of fraud against the recipient, including viruses which infect the recipient's computer to gain control over it or to return proprietary information, phishing attacks (which attempt to trick users into revealing financial passwords or PIN numbers), illegal financial schemes, and offers for products of dubious quality or legality. There are different methods some of them are as:

1.1 Reactive methods

Currently, most methods to reduce the amount of UCE in a user's inbox rely on some kind of filtering. The simple most approach is probably blacklisting, i. e. each incoming request's IP-address on a SMTP-server is tested against a list of known spamming hosts. Although, when invented back in the late 1990s, it supported the demand of switching off so called open- relays, it often has heavy side-effects: Almost all big email-providers have already been blacklisted on at least some of the widely available blacklists [3][4].

Other solutions are content-filters applied to the header and / or the body of a mail message. Filtering is based on a "bad-word-list". Later improvements include scoring- mechanisms to weight words. Those filters require a lot of fine-tuning and maintenance: Spammers are reported to register mail accounts with online services known to have spam filtering and to test their spam against those filters. This leads to a permanent "one-step-behind"-situation for filters, no matter how advanced content- filtering becomes [5].

Another still reactive way to reduce spam is grey listing, i. e. forcing the sending MTA of a message to resend it after a short time. As of now, this solution is quite potent, as most spam is sent through so called zombies, usually Windows-PCs infected with some worms. Those worms contain their own SMTP-engine, which is usually quite simple. Most of them are still unable to handle the temporary unavailable condition used in grey listing and therefore consider this condition as an error and stop delivery. However, grey listing has two major disadvantages: It slows email communication down and it is likely to be useless as soon as worms will implement better SMTP-engines, which is to be expected soon.

1.2 Modifying SMTP

The disadvantages of reactive anti-spam-methods discussed above brought the discussion on fixing one of the real causes for spam: SMTP lacks authentication. This offers spammers the chance to remain hidden and to evade lawsuits. So the key approach is to implement some kind of authentication and authorization. Beside some side-effects seen on current methods, like breaking intended mail-forwarders, the real problem is to enforce the modified standards world-wide. Beside competing standards, there are two major obstacles to handle: One is companies trying to win their share of the market by patenting their proposal for a standard. The other is due to the broad adoption of SMTP on the internet. There are billions of SMTP-clients and far more than 25.000.000 SMTP-servers in the internet [6]. Back at ARPANET times it was possible to change the standard to IP almost over night, but the internet has grown. There are still hundreds of thousands of open relays out there1, although open relays are deprecated and have been blacklisted since at least ten years. Considering this, any change to SMTP would need at least another ten years to be broadly available.

1.3 Preventing harvesters

Therefore quick-acting, effective and lasting solutions are required. A promising approach is to prevent spammers from collecting email addresses, because spammers currently only use two relevant ways to collect addresses: One is by installing worms and Trojans on computers and have them read local address books, emails or even all files, collect email addresses found there and spam to them. There is an obvious and simple solution to this: Have users install decent and safe operating systems, virus scanners and personal firewalls and protect their PCs with external firewalls and application level malware filters.

The other source of email addresses spammers use is the internet, most notably the web and the Usenet. There, they collect email addresses using spidering technology known from search engines. The programmes doing this job are called “harvesters”.

1.4 Obfuscation

Again there are some ways how to block them: One is to obfuscate email addresses, so they would not be recognised by harvesters. In [7] some different solutions are suggested, that are both compatible to any installed browser and barrier free, and proved their effectiveness in a still ongoing real world experiment [8]. Even an automated solution to dynamically obfuscate email addresses published on the web, thereby solving the problem to modify or redo existing WebPages, has been proposed [9]. Obfuscation is effective as long as the victim's address has been previously unknown to spammers. If the obfuscation method is selected with care, the email address remains human readable but is unreadable to harvesters. Therefore on well obfuscated addresses, no spam has been received yet, although the addresses were published in late 2004 [8].

1.5 Tar pit

The other approach to bar harvesters from collecting mail addresses is to trap them in a tar pit. The basic concept is to create random WebPages containing links to the same or other tar pits. This pollutes the list of webpages-to-visit the harvester has, and keeps the harvester returning and finally staying in the tar pit. As soon as the harvester is caught, all of its resources are attracted to the tar pit, thereby preventing it to visit any other webpage and collect email addresses there [10][11][6][12].

2. REMOVE AN EMAIL-ADDRESS

There are some efficient, compatible, standard-conform and barrier free ways to obfuscate email addresses. Email obfuscation has to be considered effective, but it has only been tested to work as long as the address to protect has not been published before [7]. [13] Suggested that later obfuscation of an address might also reduce the amount of spam received.

2.1 Frequency of email-address changes

The basic assumption is that people will change their email addresses quite regularly. This is certainly true for business or university environments: Most people only stay for a certain time within a certain company, therefore, they will only use their company (or university) email address for this time. According to the German institute for labour market and occupational research in 2001 28.8% of all working Germans changed into a new job [14]. This is in good accordance to the Swiss labour market: In the early 1990s each year approximately 20% of the Swiss workforce changed their job [15]. The German Federal Statistics Office said, in early 2003, 12, 8% of all employed were temporarily employed [16]. German legislation prohibits temporary employment for more than two years. Considering this, again, a fluctuation rate of approximately 25% each year sounds plausible. [17] States that the average American will change career five to seven times before retirement. This would imply a change of job approximately every six years, resulting on a roughly estimated fluctuation rate of 17%. According to [18], 1998 one out of five jobs is changed each year, reflecting a fluctuation rate of 20%. This is in good accordance to the estimation. For private email usage, those statistics do not apply. However, there are some hints to base estimations on: German phone book publishers claim in their advertising that 33% of all records would change within one year. The use of email addresses is somewhat similar to the use of phone numbers. One could bring forward the argument that email addresses do not change upon relocation. But phone numbers are now portable within a local telephone network, so they would not change for intracity relocation either. And – to compare the figures: According to [19] in 2004 11% of all Germans relocated during the year, but 33% of all phone book records changed. Therefore it sounds plausible, to presume an email address's lifetime to be approximately three years. [20] Supports this estimation by stating that an email address is valid for two years in the United States on average and within the rest of the world for three years.

2.2 Experiment presumptions

Considering the regular change of email addresses, spammers should regularly re-harvest email addresses to keep their databases up to date and add new addresses. [13] Suggests, that spammers would remove email-addresses not published any more. [20] Offers another approach by claiming in its advertising, that spammers would react to an error message within the SMTP- dialogue by removing that specific address from their database, if the bulk mailer does not use an open relay. Not to use open relays is to be considered standard among professional spammers, as blocking open relays has become a de-facto standard in spam-filtering. The Bandwidth is measures using iPerf on 60mbps and average of data is collected.

2.3 Web pages

To test the first assumption, email addresses used for the test described in [8] have been totally or partially removed from the related web page. Total removal means that all occurrences of the email addresses have been removed from the page. Partial removal means that all but one occurrence have been removed. To replace each removed address, a somewhat similar looking address has been published instead. This has been done to avoid that human visitors would notice a change on the web site and to learn how fast spammers would change to the new address and adopt it. Thanks to the tests run previously, mails sent to any address in those domains were automatically stored in a database. Therefore, the spam frequency and other data were also available for the time before this experiment started. The addresses spammed to were published for 10 month before their removal. Thereby a certain level of spam on each of those addresses has been reached.

To be able to compare data, a set of well spammed to addresses was kept on the web site. That address will be referenced as “comparison-address” here. On average, in Nov-09, each day 9.0 spam mails were received on each of those addresses. The removed addresses scored an average of 7.5 and 6.5 spam mails per day and address. On Mar-10 the addresses have been replaced. As soon as a few days later, those newly published addresses received their first spam mails.

Month / Year	Spam (compared)	removed 1	new 1	removed 2	new 2
Nov-09	271	226		200	
Dec-09	166	155		131	
Jan-10	110	128		98	
Feb-10	145	120		126	
Mar-10	148	66	13	123	16
Apr-10	323	112	20	264	11
May-10	131	159	12	189	7
Jun-10	153	145	43	174	15
Jul-10	410	184	266	305	21
Aug-10	472	228	344	331	45
Sep-10	444	216	312	222	73

Table 1: Average amount of spam mails received per address in a month.

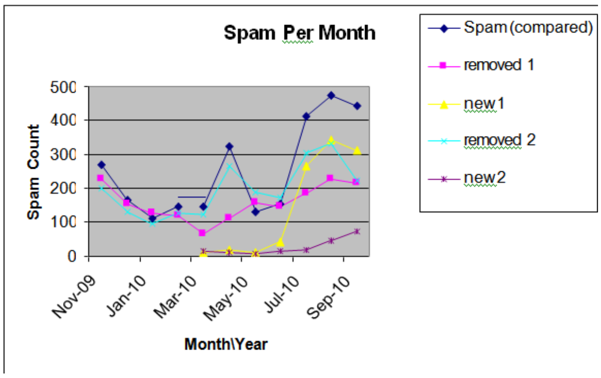
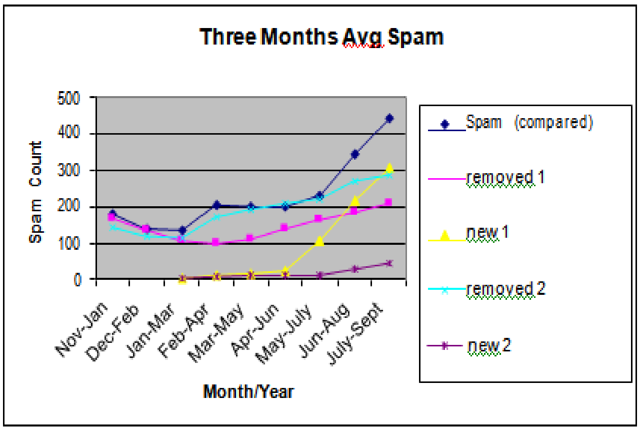


Figure -1 Graphical Representation of Spam Count

Removed 1 are totally removed addresses, removed 2 partially removed ones. Although spam totals rose in Aug -10 on the comparison-address to an average of 15.2 messages a day, on the totally removed addresses only 7.3 messages a day were received (Table 1, Figure 1). Partial address removal was not so effective, the amount of spam received kept up almost to the comparison- addresses. Only a small reduction was visible. The same effect was observed for February until May, were the amount of spam received on the comparison-address grew compared to the partially removed addresses.

This trend is confirmed by the three month average (, Figure 2). This longer term average removes some of the spam stochastic thereby increasing the reliability of the figures. It underlines that also a partial removal might have some effect as Figure 2 shows: Although the total amount of spam received on a not obfuscated address increased quickly, the three month average of the partially removed address grew much slower. The figure show the bandwidth comparison and improvement on Monthly basis and figure 4 represent the 3 months basis results of bandwidth measures be iPerf. Those figures suggest that the amount of spam received on a particular address might effectively be reduced if an email address is no longer published on a web page. Only total removal seems to be effective.



Total address removal is equivalent to effective obfuscation of email addresses. Therefore, later obfuscation of an email address is an effective anti-spam measure and helps to reduce the spam load on this address. However, it does not – at least not on the short term – bring the spam level to zero. Further research is into testing how long an email address needs to be removed from the web to be totally removed from spammer's databases. This will take some time, as spammers not only rely on harvesting to collect email addresses, but also trade them. Therefore "older" harvesting results are still available for sale in the internet. This makes removal of those addresses somewhat slower.

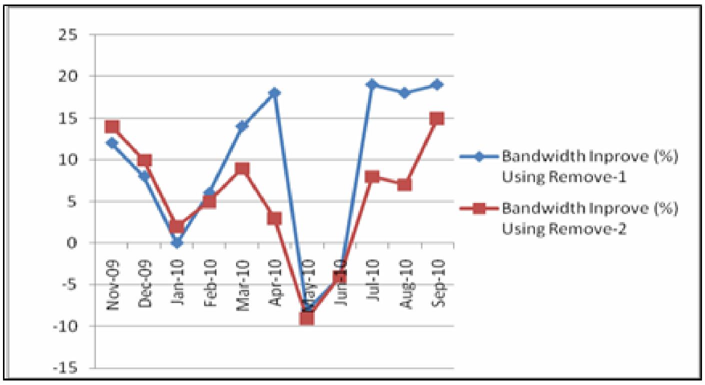


Figure 3 Bandwidth Imrovement Comparison Monthly Average

As an additional result, estimation on the speed of harvesting might be given: Within only four month, the amount of spam received on a previously unused and not published address rose to almost the same level as for an address being online for more than 18 month. Again, differences might be explained by address-trading, which probably only started for the newly published address. Address trading is also a likely reason why spam was reduced but did not totally stop on the removed addresses. It is likely that the previously harvested addresses are still available on address-CDs.

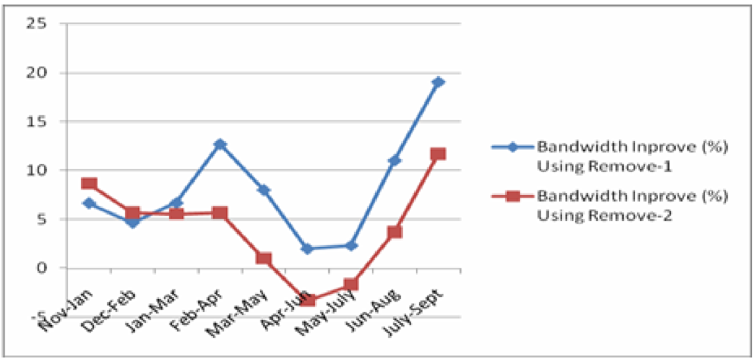


Figure 4 Bandwidth Imrovement Comparison Three Months Average

3. CONCLUSION

The Method of removing an email address from web pages, has some effect in having an email address removed from spammers' lists. In a real world environment, the existing “spots”-appliance's Basic concept is to first finish the SMTP dialogue and, if the message is considered to be spam, to temporarily reject it. The appliance stores the sender-imp, the sender's email address and the to-address. If the spammer later tries to resend this message, it is recognised by these three pieces of information. The spam-filtering mechanism would not come into effect, as there is no message to be identified as spam. Unfortunately, removing email-addresses from web pages is not as easy as it sounds, as often enough email addresses are being published out of control of the email address's owner. However, there are working solutions and there are efficient ways to obfuscate email addresses on web pages. As obfuscation proofed its effectiveness, removing an address from a web page is equivalent to obfuscating it. But even removing an email address from some parts of the internet could reduce spam, as the three month average of the partially removed address shows. To obfuscate email addresses, efficient and comfortable solutions exist, there is even an automatic way to dynamically obfuscate email addresses short before the web server delivers the content to the client. In the Mean time an average performance of 12-14 % is observed in the bandwidth utilization using iPerf which show the additional overload on the Internet and network has been reduced using this method.

References:

[1] Gaudin, Sharon, Record Broken: 82% of U.S. Email is Spam http://itmanagement.earthweb.com/secu/ article.php /3349 921, 2004.

[2] McGann, Rob, The Deadly Duo: Spam and Viruses, http://www.clickz.com/stats/sectors/email/article.php /348

3541, 2005.

[3] McWilliams, Brian, AOL lands on spam blacklist, Sebastopol, http://spamkings.oreilly.com/archives /2005/04/ aol_lands_on_sp.html, 2005.

[4] McWilliams, Brian, SpamCop blocking some Gmail servers, Sebastopol, http://spamkings.oreilly.com/archives/2006/01/, 2006.

[5] Gansterer, Wilfried et. al., Anti-spam methods - state of the art, Institute of Distributed and Multimedia Systems, University of Vienna, 2005.

[6] Eggendorfer, Tobias, Comparing SMTP and HTTP tar pits in their efficiency as an anti-spam-measure, Proceedings of Spam Conference 2006, Cambridge, MA, 2006.

[7] Gupta, Munish, Paramjeet Singh, and Shveta Rani. "Cross Layer Energy Efficient Protocols For Wireless Sensor Networks: A Survey.",*Apeejay Journal of Computer Science and Applications"*, Vol. 1, 2013, pp 27-32.

[8] Eggendorfer, Tobias, Spam proof homepage design. Methods and results of an ongoing study, Proceedings of ApacheCon 2005, Stuttgart, 2005

[9] Eggendorfer, Tobias; Keller, Jörg, Preventing Spam by Dynamically Obfuscating Email-Addresses, Proceedings of IASTED CNIS 2005, Phoenix, AZ

[10] Eggendorfer, Tobias, Ernte - nein danke. E-Mail- Adressenjägern auf Webseiten eine Falle stellen (in German: Harvesting: No thanks. How to trap harvester on a web page) in: Linux Magazin, Linux New Media, München, 2004

[11] Eggendorfer, Tobias, Stopping Spammers' Harvesters using a HTTP tar pit, Proceedings of AUUG 2005, Sydney, 2005

[12] Eggendorfer, Tobias; Keller, Jörg, Combining SMTP and HTTP tar pits to proactively reduce spam, Proceedings of SAM 2006 , SAM 2006 (The 2006 World Congress in Computer Science, Computer Engineering, and Applied Computing), Las Vegas, NV, 2006

[13] Center for Democracy and Technology, Why am I getting all this spam?, Washington, D.C., 2003, http://www.cdt.org/speech/spam/030319spamreport.pdf

[14] Schuh, Scott; Triest, Robert K., Job reallocation and the business cycle: New facts for an old debate, Federal Reserve Bank of Boston, Boston, 1998

[15] Pronto Business Media GmbH, Einwohnermeldeämter werden bürgernäher (in German: Registration offices to become more citizen friendly), Pronto Business Media GmbH, Bad Tölz, 2005

[16] IKU AG, Sponts / UCE: Nachhaltige Spam-Abwehr (in German: Lasting spam defence), Saarbrücken, http://www.sponts.de/uce.jsp, 2006

[17] Klensin, John (Editor), RFC2821: Simple Mail Transfer Protocol, o. A., 2001, http://www.ietf.org/rfc/rfc2821.txt

[18] Eggendorfer, Tobias, Spezialfilter. Antispam- Appliance mit Langzeitwirkung (in German: Special filter: Anti spam appliance with long term effects) in: Linux Magazin 09/2004, Linux New Media, München, 2004

[19] B.-A. Yassour, M. Ben-Yehuda, and O.Wasserman. Direct device as-signment for untrusted fully-virtualized, virtual machines. IBM Research Report H-0263, IBM Research Labs, 2008

[20] NLANR/DAST. Iperf. Available at http://sourceforge.net/projects/iperf.

[21] S. Bhandarkar, S. Jain, and A. L. N. Reddy. LTCP: improving the per-formance of TCP in highspeed networks. SIGCOMM Comput. Commun. Rev., 36(1):41{50, 2006}