

Spam Filtering Techniques

Daniel Owen

Middle Tennessee State University

### Abstract

This paper looks at the major spam filtering techniques in current use. In looking at methods both success rates and possible problems with each method are explored.

Methods discussed include key word filtering, open relay filtering, open proxy filtering, dial-up filtering, non conforming mailing list filtering, cooperative sharing of spam samples, known spam origin filtering and Bayesian filtering.

The first ever spam message was sent on March 5, 1994 (Moody, 2004). In the last 11 years spam has expanded to comprise approximately 65% of all e-mail (“Filtering Technologies in Symantec Brightmail AntiSpam 6.0,” 2004). As spam becomes more prevalent it threatens to make e-mail unusable. With this in mind, this paper will review several different approaches to spam filtering. Special attention will be paid to how these different types of filters operate, how they collect data and problems that the filters themselves can present.

In the last 11 years spam has expanded in magnitude from a problem that shocked Usenet users who could not believe that someone would be so crass as to advertise on the Internet but didn’t hinder normal Internet communications to a problem significantly pervasive that national governments are trying to find a way to stop, or at least limit, the amount of spam received by Internet users. To get an idea of why spam is so despised by average e-mail users and systems administrators alike you must look at the amount of spam that is sent on a daily basis. Every day AOL filters 2.4 billion spam messages. That translates to blocking 70 e-mails per user per day (Vaughan-Nichols, 2003). As an example of how bad things can easily get if spam is not curtailed consider, there are 24 million small businesses in the United States. If 1% of these companies got your e-mail address and send one message per year you would have an increase of 657 extra e-mails every day (Schwartz, 2003).

Beyond the annoyance factor, there is a cost to the spam recipient. This cost can be either in lost productivity or the monetary cost of filtering spam. Assuming that an employee can accurately delete all spam in thirty seconds per day a company with 10,000

employees can expect to spend \$675,000 per year on spam deletion (“The State of Spam,” 2003). Home users do not get off without a high monetary cost. AOL reports that they spend 15% of their users’ monthly fees on fighting spam and responding to complaints (Gaspar & Gaudin, 2001). It is obvious that spam cost a substantial amount of money for the recipient yet the cost to the sender is minimal. Many anti-spam advocates go so far as to say that spam is the equivalent of postage due advertising since the largest part of the cost is born by the recipient not the sender.

One final consideration for why e-mail is a problem is that much of what is sold is offensive or fraudulent. While there has not yet been a reported case of a company being sued because an employee received offensive spam e-mail many human resource managers worry that this could happen. Since people sending spam e-mail know nothing about their recipients it is not uncommon for children to be the recipient of explicit e-mail. This is obviously a concern for many parents. Finally, much spam advertises fraudulent merchandise. According to the Federal Trade Commission two-thirds of all spam contains deceptive or false text (Cox & Dyrness, 2003).

Considering the problems with spam it is not surprising that numerous different techniques have been developed to automate the filtering and deletion of spam. These techniques each have their relative strengths and weaknesses.

The oldest method of filtering spam is to use a blacklist. Blacklists are static lists made up of people, words or groups that have a high probability of being spam. At the simplest level a blacklist can be a list of specific e-mail addresses set up in an end user’s mail program.

The simplest form of blacklists is the word list. The idea is that certain words should never show up in legitimate e-mail so any e-mail that contains one of those words must be spam. This type of filter is typically deployed at the single user or at most the single domain level. The choice of words and phrases is extremely important in this type of filtering because almost any word can conceivably eventually end up in a legitimate e-mail. In my experience, the most effective methods of using key word filters is filtering on domain names, e-mail addresses and carefully selected phrases found in existing spam. Due to the high amount of precision that must be exercised in creating rules, this type of filter has a tendency to have high levels of false positives. One reason for the high false positive rate is that a single use of a “bad” word can get a message that otherwise looks completely innocent blocked.

In the early Internet it was not uncommon for e-mail administrators to allow anyone to send e-mail to anyone else regardless of whether either person had an account on the server that was relaying the message. This behavior is the definition of an open relay (“Open Relay Database FAQ,” 2004). Open Relay servers are a problem because some of the least scrupulous senders of spam use them as a way to hide their tracks and offload most of the cost of sending their messages to a third party. Spam operators that try to maintain a façade of legitimacy typically avoid using open relays. There are a couple of reasons for this. The use of an open relay destroys any hope of seeming legitimate and secondly it’s hard to claim that use of an open relay is not criminal computer trespass.

Blocking of open relays has certain advantages and disadvantages. Blocking open relays will cut down the amount of spam received proportionally to the amount of spam

that is funneled through vulnerable systems. Unfortunately some legitimate e-mail may also be blocked if a legitimate correspondent is using an Internet Service Provider (ISP) that has not properly secured their e-mail server. In today's environment responsible system administrators are very quick to fix any misconfiguration that might leave their servers exposed as an open relay therefore the amount of legitimate e-mail blocked should be minimal.

Open proxy blacklists are somewhat similar to open relay blacklists in that they try to stop spam operators that target misconfigured servers. An open proxy allows a spammer to send e-mail through a mail server that they would typically not have access to by making them appear to the mail server as if they were a local user (Farmer, 2003). Open proxy blacklists have similar advantages and disadvantages to open relay filtering.

Dial-up blacklists are lists that are designed to block any traffic that comes from a network address that corresponds to a consumer oriented ISP. These may be actual dial up accounts or high speed Internet accounts. The idea behind this type of list is that people in these networks should not be sending e-mail directly to other e-mail server. All e-mail should be sent through their ISP's e-mail server. Therefore there should not be any harm in blocking e-mail traffic from the portions of these networks assigned to end users. A great deal of spam has been sent using consumer ISP services through the years, so this does seem like a logical approach. Some of these messages are sent when spam mailing companies sign up for "throw away" Internet accounts. A relatively recent twist in the spam story is that some spam mailing companies have begun to hire virus writers to create viruses that allow them to send e-mail through infected home computers that act as either open relays or open proxies (Leyden, 2004). These infected computers are another

reason for spam to come from these parts of the Internet that should not typically contain servers.

Consumer oriented ISPs have been estimated to account for between 30% and 80% of all spam being sent (Bray, 2004) today. This makes it fairly obvious that a large proportion of spam can be stopped by simply blocking anything that comes from a consumer ISP. The major problem with these lists is that some small companies of home computer enthusiasts may operate their own mail servers but use service from an ISP that is listed in one of these lists. Once again there is a false positive issue.

Some mailing lists on the Internet do not confirm the legitimacy of new subscriptions. These are typically referred to as single opt-in or non confirming mailing lists. Due to the abuse of non confirmed mailing list signups by unscrupulous mailing companies some black hole list operators consider these mailing lists spam regardless of whether there have been complaints or not (“Detailed End User Information for MAPS NML Listings,” 2004). These blacklist operators advocate double opt-in or confirmed mailing lists. The difference being that in a double opt-in list the person subscribes and then receives a message that they must reply to confirming that they really want to subscribe to the mailing list. Double opt in lists have the added advantage of keeping a malicious third party for signing someone up for numerous e-mail lists in an effort to swamp them with unwanted e-mail.

Most, but not all, companies that operate legitimate mailing list have moved to double opt-in as an effort to stay off of blacklists. The disadvantage of using this type of black hole list is that there may be some legitimate mailing list e-mails that get dropped in the process of filtering out the spam. As a general rule a false positive on a mailing list

is considered less serious than a false positive on a personal e-mail but they are still a problem.

A method of filtering spam that is beginning to pick up popularity is cooperative sharing of spam signatures. This technique is similar to the method used by virus scanners in that a sample of a spam message is used to create a hash of the message. Unlike virus scanners the hash creation is automated as opposed to being a task undertaken by a human. Also unlike virus scanners all or most of the messages is used for hash creation while virus scanners typically rely on finding unique signatures within virus programs. After a sufficient number of people report the message as spam future recipients of the message will be able to automatically filter the message (Mertz, 2002).

This method is by definition more reactive than some of the other systems for spam filtering in that it relies on several people receiving and reporting the same piece of spam before it will be filtered. There is a similar problem inherent in signature based virus scanners in that they can not stop a new piece of malicious software until they have seen samples to create signatures from. In theory there should be a near zero false positive rate because e-mail must be reported by multiple people and your legitimate e-mail should be impossible to report since only you receive it. False positives can slip into the system in three ways. People forget that they are subscribed to mailing lists and report them as spam. Secondly current implementations of this method allow system administrators to configure their other spam filters to send a copy of any e-mail that appears to be spam to the central server. This means that if a mailing list gets incorrectly identified by other filters it may be reported to the central server as well. Finally it is possible, although highly unlikely, that a legitimate e-mail and a spam e-mail could end

up with the same hash if the hashing algorithm creates hashes that are not perfectly unique. This would be most likely to happen if shorter hashes were used to save storage space. In researching this I did not find any examples of this type of theoretical false positive. As with mailing list filters false positives should fall into the category of mailing lists meaning that while these false positives are problematic they are less of a problem than false positives on personal correspondence.

The final type of static list I will discuss in this paper is the known spam origin blacklist. These are lists that are comprised of email originating from IPs that have previously sent spam either to a user of the system or to a decoy address (Spews.org FAQ, 2004).

The major problem with the spam origin lists is that they are not particularly effective and have one of the highest false positive rates of any spam filtering technique. As an example, in research completed by Giga Information Group Mail Abuse Prevention Systems, LLC (MAPS), a major blacklist provider, was found to successfully block only 24% of spam but more worrying there was a 34% false positive rate (Gaspar & Gaudin, 2001).

One of the reasons for the high level of false positives by MAPS and some other known spam origin lists is that a vigilante mentality can grow in the groups that operate the lists. One common approach taken by these groups is to block “spam support” organizations. What is often means in implementation is blocking an entire ISP’s network space if they cannot get the ISP to drop a spammer.

The policy of intentionally blocking innocent customers that happen to share networks space with a spammer is called overblocking. As an example of how extreme

the overblocking can be, in February of 2002 Spam Prevention Early Warning System (SPEWS) added all of Interland's 400,000 customers to their back list because Interland had not removed 100 customers that SPEWS accused of spamming (Wagner, 2002).

These techniques are effective. Many large ISPs have caved under the pressure of having their legitimate customers blocked because they were allowing a few spammers to operate using their network. While overblocking is effective for convincing ISPs to remove known spam operators from their network it also leads to very high false positive rates making these services unusable for anyone who considers false positive results to be a problem.

The other major class of spam filtering available today is based on a self learning statistical model. There are a few different statistical models that have been discussed in the academic literature but the only one that is currently in production products, that identify their filtering method, is Naïve Bayesian filtering. It is possible that there are other statistical models that are in use in proprietary closed systems but since they are by definition closed it is impossible to consider them independently.

Bayesian filtering is based on Bayes' Theorem. The common implementation assumes that all words in a given message are not related thus, the filter is intentionally naïve and is referred to as naïve Bayesian filtering. A corpus of both spam and legitimate e-mail, referred to as ham, is collected to base the filter on. The filter looks at each word in each message and by comparing the probability of that word being in a spam or a ham messages gives it a score. When looking at new messages the filter will take a sampling of scores from the message that have the highest probability of being either spam or ham words and gives the message a score indicating that the message is either ham or spam.

Properly trained naïve Bayesian filters have reported very high filtering rates with some of the lowest false positive rates seen in any spam filtering methods. One technique that is used to reduce the number of false positive results is the doubling of non-spam words. In other words a word found in a non-spam message is twice as important as the same word found in a spam message.

One crucial issue for Bayesian filtering is the training of the filter. The more e-mail the filters sees the more accurate the assumptions about words will become. The major weakness for Bayesian filtering is that it is ideally used at the individual user level instead of at the mail gateway level. Essentially, the filter is more capable of learning the quirks of a given users good and bad words than it is of learning numerous users good and bad words since different people will have different requirements for what needs to make it through the filter. Even though this is the case several products do successfully implement naïve Bayesian filtering at the gateway level even though the success rates do take a hit (Graham, 2003). The more similar the group being filtered the more likely that naïve Bayesian filters will have results similar to those of a single user. As an example, a group of doctors will be more likely to receive drug names in their regular e-mail than the population as a whole therefore if those doctors are grouped together the false positive rate at least for those typically highly spam indicative words will remain low but if you group those same doctors with the population as a whole you will see a rise in the doctor's false positive rate and the other users of the system may see a slight decrease in the effectiveness of the filters for drug related spam.

One current product called SpamProbe tries to improve on existing Bayesian filtering by looking at word pairs and the number of times that words repeat. This is

probably the future of spam filtering as spam marketers become more adept at circumventing the existing single word statistical spam filters. This approach has many of the same advantages and disadvantages inherent in naïve Bayesian filtering. The hope is that as the techniques are improved multiple word filtering will improve even further on accuracy. A disadvantage of this form of filtering is that it does take much more storage space to store all of the seen two word combinations and probabilities (Burton, 2004).

System administrators have a final option when deciding to implement a spam filter. This option is using a system that mixes the best of the different types of systems to create an overall solution for that organization's spam problem. Most commercial packages and many of the open source solutions make a mixed approach an option. The only approach that I have seen commonly implemented without the use of any other methods as backup is Bayesian filtering at the individual user's mailbox level. By mixing different approaches the administrator has the option to weight different filtering techniques with an appropriate level of trust.

The data collection method can have a significant impact on the reliability of the results returned by the filter. As such, this is an important consideration. Different methods are appropriate for different techniques.

Static blacklist filters have three major ways that they collect data. They either scan for servers, use decoy addresses or use a nomination system.

Scanning for servers works well for open relay and open proxy blacklists. Since these are both conditions created when a system administrator has incorrectly configured the system in question it is easy for the blacklisting service to scan for these servers. Actually, what the blacklisting services that use server scanning do is fairly similar to

what spammers looking for servers to exploit do. Both groups will scan large portions of IP space for any servers that are configured as open relays or open proxies. Only their motives differ.

Decoy addresses are addresses that are specifically set up to receive spam. A real user does not ever use these addresses so there should not be any legitimate e-mail going to the address. Typically these addresses will be included as hidden text inside of a web page. This allows automated programs that look for e-mail addresses to find them without regular users getting snared in the spam trap. Anyone who sends to one of these addresses is assumed to be a spammer and added to the blacklist.

Nominations can be a blessing or a curse. They give real users who receive spam an outlet to report the spam to someone who can hopefully do something about it. Unfortunately there are issues with people sometimes forgetting about subscribing to a mailing list and then later reporting it as spam. I manage a small 30,000 user double opt-in list. I have subscribed to a service through AOL so I see any messages sent to AOL users that create a spam complain. In the last few weeks I have had at least one person every week send a spam complain about the mailing list confirmation message. These are at least in theory people who a matter of a few minutes earlier had put their e-mail address into a web form asking to receive e-mail from the list they are complaining about. There are also always a few complaints every time we send a message. I believe some of this may be because people read the subject and mistake it for spam but I also feel that a great deal of these incorrect classifications of spam come from people forgetting that they signed up for the list. As this anecdotal evidence implies people may not always be the

best way choosing what is spam in a distributed system where many people may be affected if they misclassify mailing list messages as spam.

The last method of data collection is statistical analysis. In the current systems this is a file that contains the probability of every word that had previously been seen in an e-mail message. Based on these probabilities new messages are assigned a probability of being spam.

Future statistical systems will likely use groupings of two or more words. This should help to improve accuracy by looking at the writing style of spam and legitimate messages. Multiple word statistical approaches will require a much larger corpus of training messages to give the filter the ability to see as many different combinations of word groups as possible.

These approaches to spam filtering have two ways that they help fight the spam problem. One, by blocking spam end users do not have as many garbage messages to go through. Most people are not concerned with having to delete a few garbage messages but the amount of spam has reached a point where individuals manually deleting spam have to either take a productivity reduction in carefully scanning through their e-mail or they will themselves start creating false positives by accidentally deleting legitimate messages as spam. Secondly beyond removing the annoyance factor for most users as filters become more effective more ISPs will be able to filter e-mail without having to worry about stopping their client's legitimate e-mail. As messages are tagged or deleted before they reach the end user it will be more difficult for spam senders to get their message through to the very small minority that actually buy their products. This will

lead to higher costs of operation and lower profits. If the filters are successful enough they may even remove the profit motive completely.

Spam has made e-mail less useful and more expensive and the problem is only getting worse. We currently stand at a 65% spam rate and the amount of spam doubles every 12-18 months. Now is the time for e-mail filtering to become prevalent. There are numerous different methods of filtering and the goals of the person doing the filtering will to a great extent determine which method they chose to employ. The question is no longer whether to filter spam but what method to use in filtering e-mail.

## References

- Bray, H. (2004, June 9). Home PCs big source of spam. Retrieved November 16, 2004, from [http://www.boston.com/business/technology/articles/2004/06/09/home\\_pcs\\_big\\_source\\_of\\_spam/](http://www.boston.com/business/technology/articles/2004/06/09/home_pcs_big_source_of_spam/)
- Burton, B. (2004). SpamProbe - Bayesian spam filtering tweaks. Retrieved October 17, 2004, from <http://spamprobe.sourceforge.net/paper.html>.
- Cox, J. & Dyrness C. (2003, May 28). Spam prevention may lead to filtering of legitimate messages [Electronic Version]. Knight Ridder Tribune Business News, 1.
- Detailed end user information for MAPS NML listings. (n.d.). Retrieved November 16, 2004, from [http://www.mail-abuse.com/support/enduserinfo\\_nml.html](http://www.mail-abuse.com/support/enduserinfo_nml.html).
- Farmer, J. (2003, December 27). An FAQ for news.admin.net-abuse.email part 3: understanding NANAE. Retrieved October 3, 2004, from <http://www.spamfaq.net/terminology.shtml>.
- Filtering technologies in Symantec Brightmail AntiSpam 6.0. (n.d.). Retrieved November 15, 2004, from <https://enterprisesecurity.symantec.com/content/displaypdf.cfm?SSL=YES&PDFID=1025>.
- Gaspar, S. & Gaudin, S. (2001, September 10). Spam police. Network World, 18(37), 58-62.
- Graham, P. (2003, January). Better Bayesian filtering. Retrieved October 17, 2004, from <http://www.paulgraham.com/better.html>.
- Leyden, J. (2004, May 14). Spam fighters infiltrate spam clubs. Retrieved November 16, 2004, from [http://www.theregister.co.uk/2004/05/14/spam\\_club/](http://www.theregister.co.uk/2004/05/14/spam_club/).
- Mertz, D. (2002, August). Spam filtering techniques: Comparing a half-dozen approaches to eliminating unwanted email. Retrieved November 16, 2004, from <http://gnosis.cx/publish/programming/filtering-spam.html>.

- Moody, G. (2004). Spam's tenth birthday today. Retrieved November 16, 2004, from [http://news.netcraft.com/archives/2004/03/05/spams\\_tenth\\_birthday\\_today.html](http://news.netcraft.com/archives/2004/03/05/spams_tenth_birthday_today.html)
- Open relay database FAQ. (n.d.). Retrieved November 16, 2004, from <http://www.ordb.org/faq/>.
- Schwartz, E. (2003, July/August). Spam wars. *Technology Review*, 106(6), 32-39.
- Spews.org FAQ. (n.d.). Retrieved November 13, 2004, from <http://spews.org/faq.html>
- The state of spam Impact & solutions. (2003, January). Retrieved November 13, 2004, from [http://web.archive.org/web/20030621231814/www.brightmail.com/press/state\\_of\\_spam.pdf](http://web.archive.org/web/20030621231814/www.brightmail.com/press/state_of_spam.pdf).
- Vaughan-Nichols, S (2003). Saving private e-mail. *IEEE Spectrum*, 40(8), 40-44.
- Wagner, J. (2002, May 23). When spam policing gets out of control. Retrieved October 17, 2004, from [http://www.internetnews.com/xSP/article.php/8\\_1143551](http://www.internetnews.com/xSP/article.php/8_1143551).