

Featured Article

Spam and Virus Filtering

Spam, or unwanted bulk email, and viruses have become an increasing problem over the last few years. In 2003 it was estimated that corporations lost \$10 billion worth of diminished user productivity, consumption of IT resources, and help desk costs in combating spam.ⁱ The costs associated with viruses are much higher. In February 2004, viruses cost an estimated \$83 billion worldwide.ⁱⁱ The main target vector of viruses over the last several years have been through email, so dealing with these issues together is ideal. There are hundreds of plugins out there for email clients that can help with spam and virus identification, but if the spam/virus has made it to your computer then it has won half the battle. Let's examine the various ways that the spam/virus can be identified at the server level or even before the server.

Rule Based Scanning

Rule Based scanners use rules to look at email and assign weights to words or phrases which are commonly used in spam. For example some sites might want to block all email containing explicit or vulgar words. This can work, but it can also lead to some false positives. As words such as 'sexual' usually connote spam, they also can be important (non-spam), e.g. 'sexual offender'. Also, the simple word filter is hard to implement since spammers quickly catch on and obfuscate their words. Consider the word V14gra or V I a g r a. We can immediately read these words, but computers cannot. Spamassassin is the most commonly used rule based scanner and has an extensive database of rules which assign weights. For example consider the rule:

```
body DRUGS_SMEAR1 /(?:Viagra|Valium|Xanax|Soma|Cialis){2}/i
describe DRUGS_SMEAR1 Two or more drugs crammed together into one word
score DRUGS_SMEAR1 1.310 1.372 1.576 1.337
```

This rule looks at the body of the email, for two or more drug names together as one word. The score section contributes a weight to the overall message score and the sum of the weights can determine whether or not a message is spam. The scores are written so that you can adjust them to suit your needs. Alabama Supercomputer Authority's email service has been tuned so that the rules for some words are weighted much higher than the standard to catch even more spam.

Bayesian Filtering

One of the weaknesses in the rule based approach is that the rules are fixed and once you have a copy, the spammers do as well. They can then reverse engineer their spam to bypass the rules. Bayesian filtersⁱⁱⁱ adapt to the content by relying on the user to 'teach' it what is spam and what isn't. Ideally, one would start with a large corpus of known good mail and known spam. Then the filter would learn what is good for you and what is bad. Each time you get a spam message you can send it to the filter and it will adjust the scoring system depending on what it gets from you.

Whitelisting and Blacklisting

Whitelisting (always allowing) and blacklisting (always blocking) are settings in every spam filter. These are the first line of defense against viruses and spam floods, and would also enable us to receive email from your favorite uncle even if he liked to talk about all the pain medication he is on after his surgery. Now certain sites publish their own blacklists of known spammers, called Real-time Black Lists (RBLs).

RBLs^{iv} are lists of IP addresses that spam is known to have originated from. When applied to your mail filter now, you can eliminate a good quantity of email, and depending on your configuration keep the email from reaching your server. False Positives (non-spam marked as spam) are a large problem with RBLs. For instance, if someone gets a virus on a private network, the public address is listed with the RBL, and no one behind the firewall can send mail. Also, legitimate mailservers occasionally are listed on RBLs due to a miscommunication, or misconfiguration. Recently this has happened to several big names such as Bellsouth and Earthlink. This is a major concern. The best way is to use the RBL is in conjunction with a rule based scanner so that weights for each RBL can be established, and contribute to the overall spam score. ASA has its own RBL list which is constantly being updated to augment our spam and virus scanners. This list is publicly addressable and we may be interested in helping others by publishing a global whitelist and blacklist in the future.

Checksum Scanning

Blocking unsolicited bulk email can also be facilitated by noticing the word 'bulk'. Since the vast majority of spam is actually 1 message sent to thousands of people, the best way to block it is simply by counting each message seen by a mailserver. A unique tag or 'checksum' of each message is used to count the number of messages. These checksums can be shared to give a 'global' count. Once the checksums have been seen many times, and the IP address of the originating machine has not been whitelisted, then the emails are marked as spam. Distributed Checksum Clearinghouse^v has a network of servers that simply count each unique email, and then share that information with all of their partners. As the number of times that each email has been seen rises, its probability of being spam or a virus also rises. ASA is currently a member of the DCC network of servers which drastically reduces the timeouts and other errors associated with network checksumming.

Another checksumming system is that of Razor^{vi} or Pyzor^{vii}. Each of these has a network of actual people who report known spam checksums. In this way, it is not the quantity of email that is important; simply that one of their network of people has reported a message as spam.

Sender Policy Framework

Sender Policy Framework^{viii} (SPF) actually has very little to do with determining what is spam. What SPF attempts to do is to increase the 'cost' of spam to the sender. Currently, spammers pay very little for the messages they get out. But if one could increase the cost to them then this would be a disincentive to spam. SPF does this by

registering which machines in your domain are registered to send mail for your domain. In this way, forging "From:" headers in the email message become harder. If everyone eventually registers an SPF record then spammers will be easily identifiable and finding their domains and blocking them becomes quite simple. Currently AOL and Yahoo are already implementing SPF checks on all incoming messages, with hard failures to be implemented soon. For ASA customers who have their DNS registered with our nameservers, we offer SPF registration as a free service, simply call the helpdesk and tell them you would like to register your SPF record in DNS.

Virus Scanning

This is the last item on our list but it is probably the most necessary. Email is the method of choice for distributing viruses. Running a spam filter without catching the viruses only fixes part of the problem. The most popular thing for viruses to do (other than replicate themselves) is to send out more spam. So implementing a virus scanner as part of your overall spam solution is an absolute necessity.

Off Site Scanning

Your spam and virus scanner does not have to be on your network. If you consider that the spam and viruses are just wasting your bandwidth then you probably would not even want the scanner on your network. Solutions exist whereby a simple DNS change can direct incoming email to an external mail scanner through a 3rd party provider. **ASA provides such a service today to many school systems.** ASA's email scanners implement all of the rule based systems above, with some custom tweaks which enable us to catch even more of the spam and viruses. Further restrictions can be applied which will deny any messages that are not from ASA's mailscanners to your machine.

Management of ASA's mailscanner service is through a web-based interface in which you can see what messages have been filtered. You have the ability to release at least some of them from the quarantine if they are falsely identified as spam. This interface also gives you statistics on how much email, both volume and quantity, the filter is blocking for you. Currently ASA averages between 75% and 80% of all email currently being rejected as spam. **ASA is also keeping over 2 GB of email per day off of our customers' Internet connections.**

The screenshot shows the MailWatch web interface. At the top, there are several summary boxes:

- Color Codes:** Bad Content Detected (red), Spam (orange), High Spam (yellow), MCP (green), High MCP (blue), Whitelisted (purple), Blacklisted (grey), Clean (white).
- Status:** MailScanner: 75.323 2.5GB, Prefs: 1.133 1 proc(s), Load Average: 0.21 0.08 0.02.
- Today's Totals:** Processed: 71,323 2.5GB, Clean: 24,684 34.6%, Viruses: 3 0.0%, Top Virus: Wurm_Smurf.P, Blocked Files: 48 0.1%, Others: 1 0.0%, Spam: 17,619 23.9%, High Scoring Spam: 29,548 41.4%, MCP: 0 0.0%, High Scoring MCP: 0 0.0%.

Below these is a table titled "Last 50 Messages (Refreshing every 30 seconds)":

#	Date/Time	From	To	Subject	Size	SA Score	Status
[1]	14/03/06 14:00:14	phw@eth@bnetmail.net	scans@hsv.k12.il.us	A Good Laugh on Friday	3.4KB	0.88	Clean
[1]	14/03/06 14:00:14	cary@rcisul.rcisul.wednet.edu	jag@calhoun.edu	RE: Service #360737	1.6KB	0.00	Clean
[1]	14/03/06 14:00:14	0higgins@teacher.com	df@calhoun.edu	Your Order canceled, contact us	9.1KB	0.00	Spam
[1]	14/03/06 14:00:14	sanosurvey@postac.org	mery@hsv.k12.il.us	The survey code you requested	2KB	4.90	Clean
[1]	14/03/06 14:00:14	enb@hsv.k12.il.us	enb@hsv.k12.il.us	Your new cell phone service is pending activation	4.6KB	30.77	Spam
[1]	14/03/06 14:00:13	us.din@arc@gangway.ecopy.net	enram@hsv.k12.il.us	Your Hydrolex Sample - Please respond	2.9KB	9.87	Spam
[1]	14/03/06 14:00:12	access@dfhangar.com	hsvkino@hsv.k12.il.us	SkyRocket Your Google Ranking	4.2KB	14.78	Spam
[1]	14/03/06 14:00:12	kate_b@netnet.com	tnberlake@trant.k12.il.us	RE: Donkey	39.4KB	1.11	Clean
[1]	14/03/06 14:00:12	mery@hsv.k12.il.us	jaberson@hsv.k12.il.us	KE: Item #7235726156 - Notification of an Instant Payment Received from jones (jaberson@hsv.k12.il.us)	18.1KB	0.11	Clean
[1]	14/03/06 14:00:11	stanada@gmail.com	enb@hsv.k12.il.us	Fw [01] April-04 (Tabloid) - The "N.g.e.r V.1"	8.3KB	32.11	Spam
[1]	14/03/06 14:00:11	carolb@hsv.k12.il.us	hamb@hsv.k12.il.us	FW: The Call Of Good Times - Do U Remember (UNCLASSIFIED)	24.6KB	0.40	Clean
[1]	14/03/06 14:00:11	2-039146-sci.k12.il.us	smoath@hsv.k12.il.us	Put Your Best Foot Forward - Complimentary Fall Color Business Cards	4.8KB	22.41	Spam
[1]	14/03/06 14:00:10	wj@asc.edu		Mail System Error - Returned Mail	28.4KB	0.91	Clean

Conclusion

There are many tools with which to fight the problem of spam. Implementing them now can save time and money for your network. Consider how much time and energy you currently spend on fielding calls related to spam and viruses. As a first step, please consider publishing a valid SPF record. This will definitely help with the overall problem in the future.

ⁱ The Ferris study, <http://www.entmag.com/news/article.asp?EditorialsID=5651>

ⁱⁱ The Washington Times on Monday, March 01, 2004 Article ID: D140718

ⁱⁱⁱ <http://spamassassin.apache.org/>

^{iv} <http://www.email-policy.com/Spam-black-lists.htm>

^v <http://www.rhyolite.com/anti-spam/dcc/>

^{vi} <http://razor.sourceforge.net/>

^{vii} <http://pyzor.sourceforge.net/>

^{viii} <http://www.openspf.org>

For more information please contact Richard Trice(rtrice@asc.edu).