COMPARATIVE REVIEW

ANTI-SPAM COMPARATIVE REVIEW MAY 2009

Martijn Grooten

If you happened to pass the *Virus Bulletin* office during the last few days of April, you would have been forgiven for thinking you had heard the popping of champagne corks in celebration of the completion of our first comparative anti-spam test. After months of consideration, internal and external discussion, trials and retrials, we are very pleased to be able to reveal the results of the first test.

Still, much as we believe that our test is a good one, we are the first to admit that there is room for improvement – indeed we are already working on a number of adjustments to the test set-up. Moreover, there were a couple of minor bugs that had to be fixed during the course of the test, and it is only fair that we confess to these issues.

One of the things that went wrong was that, one week into the test, the primary DNS server failed. Most products use a secondary DNS server as a backup solution, as do our own servers, and it was for this reason that we did not notice the problem until later on. The problem was brought to light when one of the products on test showed a significant drop in performance – it turned out that the product in question was only using the primary server for DNS lookups. While it is generally assumed to be best practice for products to use at least two DNS servers, this requirement had not been stipulated prior to the start of the test – we intend to make this a formal requirement for entrants in future tests. Of course, we have also learned that it is important to monitor the performance of the DNS servers closely.

A second bug was caused by a minor error in the script used to relay email to the products. This resulted in some of the emails being relayed incorrectly. Thankfully, a comprehensive logging system meant that we were able to identify these emails easily and, after fixing the bug, remove them from the test set.

THE TEST CORPUS

The test corpus consisted of all emails sent to the virusbtn.com domain between the afternoon of 9 April and the morning of 30 April 2009. The original idea was to let products filter all email, regardless of whether they were sent to an existing address, thus maximizing the amount of spam seen by the products. However, not all of the products could be configured in this way and as a result we decided to remove from the corpus any messages that had been sent to addresses that do not correspond to a genuine *VB* mailbox or alias. After removing these, as well as the misrelayed messages, the test set consisted of 1,677 ham emails and 24,320 spam emails. The ham set included personal and business email, newsletters, mailing lists, genuine delivery failures and automated notifications. The nature of some of these emails (in particular automated notifications, newsletters and mailing lists) makes them very difficult to distinguish from spam. Nevertheless, they are all messages that the virusbtn.com end-users genuinely want to receive, and as such they should not be blocked by a spam filter. It should be noted, however, that the false positive (FP) rates recorded in this test may be higher than those reported in other tests using ‘easier’ ham corpora (containing fewer newsletters, mailing lists and so on). This is one of the reasons why the absolute numbers shown in the test results do not give a good picture in isolation; it is the relative numbers compared to those of other products that demonstrate how well a product performs.

To determine the ‘golden standard’ for each email, we first applied some ad hoc rules. For example, we determined that any message using a foreign alphabet was almost certainly spam. It should be noted that under the test regime, products are not allowed to make use of such ad hoc rules based on *VB*’s assumed email behaviour – and regular checks are carried out to ensure this is not the case. Secondly, if all products agreed on the classification of an email they were assumed to be correct; again, we performed regular checks to ensure that nothing was misclassified (even though the comparative nature of the test would mean that a mistake here would not disadvantage any product).

Finally, for all remaining emails, the golden standard was decided upon by the end-user – the *VB* employee to whom the email was sent (see p.2). To minimize the effect of human error, all emails reported as false positives by at least one of the products were double-checked to ensure the correct classification had been made by the end-user.

THE TEST SET-UP

A brief description of the test set-up follows below. Full details of the set-up and the thought processes behind it can be found in *VB*, January 2009 p.S1; *VB*, February 2009, p.S1 and *VB*, March 2009, p.S6.

A gateway Mail Transfer Agent (MTA) running qpsmtpd 0.40 on a *SuSE10 Linux* machine was configured to accept all email sent to the virusbtn.com domain. Upon accepting an email, the MTA stored it in a database then relayed it to all participating products in random order. The original email was unchanged with two exceptions: first, a Received header was added to reflect the fact that the email had passed through our MTA. Secondly, if the email header lacked a Message ID, one was added using the mail.virusbtn.com domain.

All of the products participating in the test were configured to relay the filtered email to a back-end MTA. Where possible, they were configured to relay spam as well, and to mark spam using a special header. Using this header in combination with the IP address on which the product was located, the back-end MTA was able to link a filtered email with both a product and an email that was already in the database.

Two of the products, *ClamAV* and *SpamAssassin*, were not installed on a server; instead they were installed on the same machine that runs the MTA. For performance reasons, emails were not sent through these two products immediately after they were received. Instead, a script checked every 10 minutes for new messages then ran them through both filters.

BitDefender Security for Mail Servers 3.0.2

SC rate: 84.20%

FP rate: 1.49%

FP of total mail corpus: 0.096%

BitDefender is no stranger to *Virus Bulletin*, since the Romanian vendor is a regular participant in the VB100 anti-malware reviews. The company has also been active in the anti-spam business for quite some time and was one of the first to submit a product for this test.

BitDefender Security for Mail Servers comes in various flavours for different operating systems; the version we tested ran on a new *SuSE10 Linux* installation as an extension (milter) to the *Postfix* MTA. Installation of the product was straightforward and consisted of downloading an executable .rpm file and running it. The product can be configured using the command line, which no doubt will please many experienced *Linux* administrators, but those who prefer a graphical interface will also find themselves at ease with the web interface.

BitDefender's false positive rate was lower than that of any of its commercial competitors. The spam catch (SC) rate, however, left some room for improvement. The low spam catch score is partly explained by the product's use of only one DNS server – something the developers have since fixed. Indeed, during the period in which our primary DNS server was down, the product's performance dropped about six per cent. Despite this, the product's performance was more than decent and, while working on improvements to the product for the next test, its developers will be able to revel in the knowledge that they have already achieved a VBSpam Gold award.

ClamAV using Sanesecurity signatures

SC rate: 27.63%

FP rate: 0.00%

FP of total mail corpus: 0.00%

ClamAV is the biggest and best-known open source anti-malware product and is developed by a large group of volunteers from all over the world. While many anti-malware reviews suggest that *ClamAV*'s performance falls short of that of its commercial competitors, it still boasts many happy users. In particular, many of them use the product on mail servers to check incoming and outgoing email for malware. However, it can also be run as a spam filter, and as such it was submitted to the test. The scanning rules were based on signatures provided by a group of volunteers operating under the name *Sanesecurity*.

We had been warned that the spam catch rate would be far from that of dedicated anti-spam products and indeed, we found that the product blocked barely 28% of all spam. However, that does not render the product worthless. The fact that, even in our difficult ham corpus, no legitimate message was blocked incorrectly indicates that the product could act as a very good first-layer filter, working in conjunction with a number of others. Moreover, the nature of signatures is such that the product's performance might change significantly if it were to see a different email corpus (indeed, we saw great variation in its day-to-day performance), and I will be very interested to see how it performs in the next test, using a larger spam corpus.

MessageStream (Giacom)

SC rate: 96.50%

FP rate: 3.16%

FP of total mail corpus: 0.204%

Giacom's *MessageStream* is a hosted solution that takes the spam filtering away from the customer's mail server: email is passed through and filtered by *MessageStream*'s servers, where spam is quarantined and only presumed ham messages are sent back to the customer's mail server.

An attractive and intuitive web interface is available for the configuration of product settings as well as for the whitelisting of email addresses or full domains on either a global or personal level. I was charmed by the information that is provided on why emails have been marked as spam – enabling users to modify filter rules even if they aren't experts on spam filtering. I was less excited by the fact that there is no facility for an administrator to search all email (sent to all addresses) simultaneously, but end-users' privacy is more important than saving the system administrator a few minutes' work.

On the company's website, the product is claimed to block at least 97% of spam and our test results indicate a similar score – far above the average. Unfortunately, there were a few false positives, but judging by the spam scores for the emails in question, most of them could probably have been avoided (albeit at the cost of a lower spam catch rate) by tuning down the spam filter slightly. A VBSpam Gold award is thus very well deserved.

M+Guardian (Messaging Architects)

SC rate: 94.83%

FP rate: 2.27%

FP of total mail corpus: 0.146%

For those companies who want to keep their anti-spam solutions in house, yet do not want to configure a server themselves, a hardware appliance might be the right choice. One of the many available on the market is *M+Guardian*, a product from Canadian company *Messaging Architects*. The appliance can be stored in a server room like any other server, with the difference that you don't have to worry about installing and maintaining an operating system.

vb May 2009 platinum SPAM virusbtn.com

Like most products, *M+Guardian* comes with an easy-to-use web interface for product configuration and the monitoring of email flow. I liked the fact that there was an option to send a warning once the number of spam emails received by a single user has exceeded a certain threshold – thus reminding end-users that using one's address sensibly is the first step to minimizing spam.

While the product did generate some false positives, the number was lower than average. Add to that a very high spam catch rate and *M+Guardian*'s developers can be proud to be the first to achieve a VBSpam Platinum award.

SpamAssassin

SC rate: 61.41%

FP rate: 1.07%

FP of total mail corpus: 0.069%

With a history dating back to 1997, *SpamAssassin* is the Methuselah among anti-spam products. The product is far from retirement though, and it is used as heavily as ever and still worked on by a large group of volunteers. Operating under an *Apache License 2.0*, the product is free and open source. For this test, we used version 3.1.8 on *SuSE10 Linux*, which was updated every hour.

I do not believe that using free anti-spam software is necessarily a better idea than using a proprietary product, nor do I think that the performance of a free product is bound to be worse than that of a commercial product. However, the vendors of commercial products need good reason to expect customers to pay for their wares if decent free alternatives are available – so it will be interesting to see how performances compare.

Unfortunately, *SpamAssassin*'s spam catch rate was a disappointingly low 61% – which was barely compensated for by a very low false positive rate. Undoubtedly *SpamAssassin*'s developers will be as curious as I am as to whether the low spam catch rate was caused by loose filter rules that need tightening, or whether other factors have also played a role.

Webroot E-Mail Security SaaS

SC rate: 97.57%

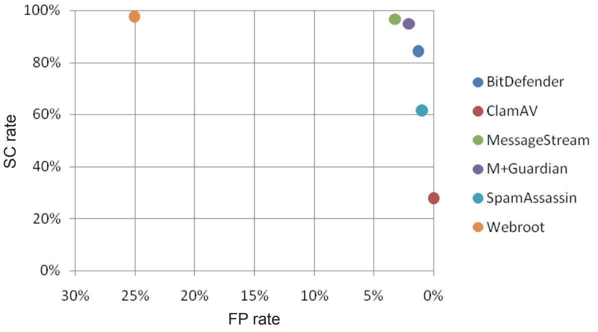
FP rate: 26.12%

FP of total mail corpus: 1.685%

Webroot is another vendor that will be familiar to *VB* readers from its participation in VB100 tests, and was another that submitted a hosted solution. Like most hosted solutions, *Webroot* does a lot more than simply filtering spam – other functions include the provision of business continuity and scanning of email for pornographic images. In an era in which more and more spam is sent from compromised legitimate machines, it is also reassuring to see that the product can be configured to scan outbound messages.

A decent web interface gives system administrators a good overview of current spam and virus threats, as well as an indication of which users are most affected. Unfortunately, due to the way in which the product was set up for this test, few of the options in the interface could be tried out.

In fact, *Webroot*'s developers are already working on finding a way to make the product fit into the test better: a false positive rate of over 25% of all ham messages is almost certainly a sign of product misconfiguration. With



	True negatives	False positives	True positives	False negatives	SC rate	FP rate	FP rate as percentage of total mail corpus
BitDefender Security for Mail Servers	1,652	25	20,478	3,842	84.20%	1.49%	0.096%
ClamAV signatures	1,677	0	6,719	17,601	27.63%	0.00%	0.000%
Giacom	1,624	53	23,470	850	96.50%	3.16%	0.204%
M+Guardian	1,639	38	23,062	1,258	94.83%	2.27%	0.146%
SpamAssassin	1,659	18	14,934	9,386	61.41%	1.07%	0.069%
Webroot E-Mail Security SaaS	1,239	438	23,729	591	97.57%	26.12%	1.685%
Average					**77.02%**	**5.68%**	**0.367%**

such a high false positive rate no certification was awarded this time around, but the developers will no doubt be working hard to achieve significantly better results in the next test.

AWARDS

It cannot be emphasized enough that, in our tests, it is not so much the *absolute* performance of a product that matters, but the *relative* performance compared to that of its competitors. Products will therefore not achieve certification by blocking ‘x%’ of all spam or generating less than ‘y%’ false positives. The best-performing products in each test are awarded with one of three certifications:

- VBSpam Platinum for products with a spam catch rate twice as high and a false positive rate twice as low as the average in the test
- VBSpam Gold for products with a spam catch rate at least as high and a false positive rate at least as low as the average in the test
- VBSpam Silver for products whose spam catch rate and false positive rates are no more than 50% worse than the average in the test.

In this test, based on an average spam catch rate of 77.02% and an average false positive rate of 5.68%, the benchmarks were as follows:

Platinum: SC 88.51%; FP 2.84%

Gold: SC 77.02%; FP 5.68%

Silver: SC 65.53%; FP 8.52%

One does not need a qualification in statistics to understand that these averages have been skewed by the performances of *ClamAV* (which had a very low spam catch rate) and *Webroot* (which had a very high false positive rate). It would thus be tempting to ignore these products when computing the average score. However, we have decided against this based on the fact that we think it is important to stick to the same rules for the duration of a test, rather than change them halfway through.

We are looking into ways in which any future ‘outliers’ can be excluded from the calculation of averages using non-arbitrary methods.

CONCLUSIONS AND IMPROVEMENTS

If there were two changes I could make to improve the test they would be:

- The inclusion of more products in the test.
- The use of a larger and more varied spam corpus.

Happily, thanks to a great deal of interest from vendors, we anticipate that the number of products participating in the next test (due to be run in June) will reach double figures.

To increase the size of the spam corpus and the variation within it, we intend to work together with *Project Honeypot* – an initiative that has generated the largest and most varied spam trap in the world. The brains behind *Project Honeypot* have kindly offered to relay some of the millions of spam messages they receive to our servers, so that they can be used in our test in real time. This will significantly increase the robustness of the test.

Overall, despite a couple of bugs the first ‘live’ anti-spam test has been a success, with some encouraging results for most of the participants and a little more work to be done by some of the others. I look forward to the next test to see the effects of a larger field of competition and a larger spam corpus.

Developers interested in submitting products for the next test should contact martijn.grooten@virusbtn.com. The next test will be run during June, with the deadline for product submissions towards the end of May.