Comparing Industry-Leading Anti-Spam Services

Head-to-Head Testing Results

Joel Snyder
Opus One®
December, 2015

INTRODUCTION

The following analysis summarizes the spam catch and false positive rates of the leading anti-spam vendors. Compiled by Opus One, an independent research firm, this report provides data to objectively compare the market's most popular anti-spam solution.

In total, nine products and services were evaluated using strictly objective criteria to measure their ability to block unwanted email with minimal false positives. The products were chosen to include the vendors with significant market share taken from Gartner's June 2015 Magic Quadrant for Secure Email Gateways, and includes products from Barracuda Networks, Cisco Systems, Intel Security (McAfee), Microsoft, Proofpoint, Sophos, Symantec, Trend Micro, and Websense. Trend's InterScan Message Security solution was evaluated and included in these test results. The remaining vendor names have been obfuscated.

TEST METHODOLOGY

Opus One has run regular monthly tests of anti-spam software since 2004. The results reported in this white paper are from our 117th test, in September 2015. To ensure consistency and reliability, Opus One uses the same methodology each month, providing the opportunity to compare performance of products over time. In this test:

- Approximately 12,000 messages were selected at random for testing
- Messages were drawn from actual corporate production mail streams
- Messages were received live and tested with less than a one-second delay
- Tested products were acquired directly from the vendor or through normal distribution channels and were under active support contracts. All products except for Office 365 were tested on-premises. Tested products were "up to date" with current released software and signature updates and were configured as recommended by the vendor's own technical support team
- Messages were hand classified as "spam" and "not spam" to ensure data validity
- Each of the tested products included the vendor-recommended or integrated reputation service in the results

Table of Contents

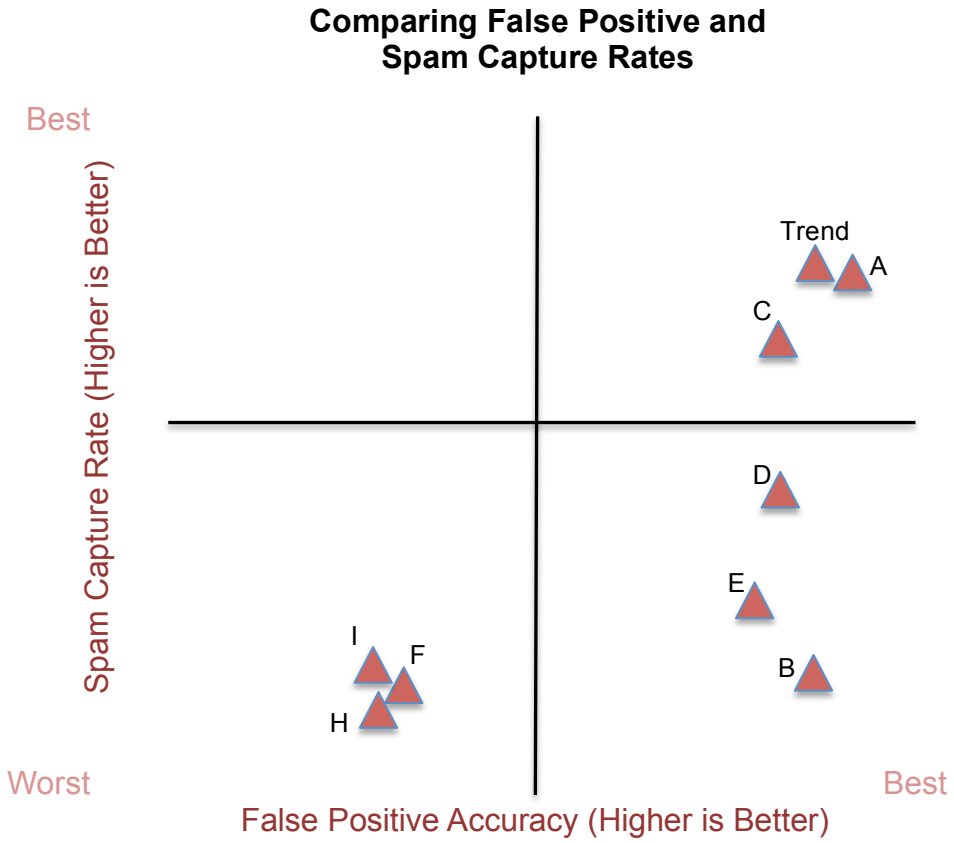
With ten years of monthly results, Opus One is uniquely positioned to provide objective efficacy reporting across all major anti-spam products. *While testing occurred in North America, message sources were global. See the appendix at the conclusion of this report for further test methodology details and definitions of terms.*

OVERVIEW OF TEST RESULTS

Trend's email security gateway was one of the products that clearly stood out for having the highest spam capture rate and the most accurate rate of detection. In the graphic below, we have placed vendors on a grid with the best performing products in the upper right quadrant. The combination of excellent catch rate and low false positive rate makes it clear that there is a significant difference in both completeness (catching most spam) and accuracy (lack of false positives) in the most popular products.

The tradeoff between false positives and spam capture rate is less obvious than it has been in past years. We would expect products to cluster either in the upper left (good spam capture, but high false positive rate) or lower right corners (good false positive rate, but lower spam capture). In fact, products D, E, and B act as expected, but products from vendors A, C, and Trend have overcome the statistical barrier and give both good capture and false positive results.

The results summarizing false positive rate and spam catch rate are summarized below.



DIFFERENTIATING SPAM CAPTURE RATES

The spam catch rate has a direct impact on end-users' satisfaction with the email system and their productivity. With the high daily global volume of spam, even the slightest reduction in catch rates can have a major adverse effect. There may not seem like a lot of difference between a catch rate of 97% and 99%, but when the number of spam is so high, small differences add up to a lot of spam delivered to users' inboxes. The amount of spam in our mail feed varies each month as spammers come and go. During the month reported here, 94% of the incoming mail feed during the test period (a total of 11,989 messages) was spam. Each of the vendors missed some, but the differences are extreme, with the worst-performing vendor delivering more than three times as much spam.

The spam catch rates below are sorted by the number of missed spam. The ordering of the products is somewhat different from the summary chart because we're simply dividing the missed spam by 11,272 (the total number of spam).

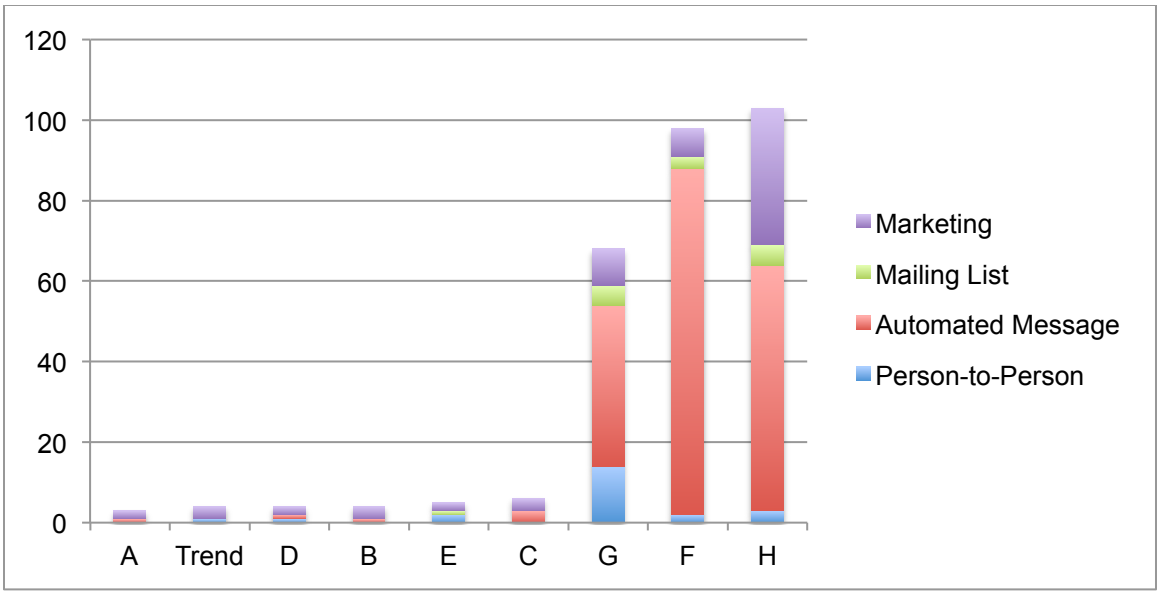
Vendor	Spam Catch Rate
Trend	99.0% (109 missed spam)
A	99.0% (111 missed spam)
C	98.6% (155 missed spam)
D	97.8% (245 missed spam)
E	97.3% (304 missed spam)
B	97.0% (335 missed spam)
I	96.9% (346 missed spam)
F	96.9% (351 missed spam)
H	96.8% (365 missed spam)

FALSE POSITIVE RATE

Because of the mission critical nature of email, it is essential that an enterprise's anti-spam solution deliver a low false positive rate. Messages incorrectly quarantined and blocked pose a serious loss of time and productivity for system administrators and end-users. Users have come to trust their anti-spam products to have vanishingly low false positive rates, but our testing shows that this confidence is unfounded: three products (I, F, and H) had significantly higher false positive rates than the rest.

Products from vendor A, D, E, and Trend all have essentially identical false positive rates—the difference between the lowest (2 false positives) and the highest (4 false positives) is within the range of statistical error of a test like this.

Our testing categorizes false positives according to the type of message: person-to-person (the worst type of false positive), automated messages (such as "your package will be delivered tomorrow"), mailing lists, and marketing messages. We believe that it is important to distinguish between products that have different types of false positives. When a marketing message doesn't get through, this is not very important; when a person-to-person message is lost, this can have a significant impact. The chart below summarizes the results sorted by category. We don't include the effects of reputation services, because false positives from reputation services are easily detectable, unlike most other false positives.



CONFIGURATION NOTES

Each anti-spam product has different configurations. In our testing, we use the product vendor's technical support team to advise us on the best settings for our environment. However, there are always setting options which are more policy-based than efficacy-based. In the results reported here, we have chosen:

- to pick default settings wherever possible;
- to use the vendor's own anti-spam engine (where a choice of engines is available);
- to use the vendor's own reputation service, if the vendor has one; if not, we have used Spamhaus reputation service;
- to use the reputation service aggressively (i.e., low threshold); and
- to ignore "suspected spam" (if the product has such an option)

ABOUT OPUS ONE

Opus One is an information technology consultancy with experience in the areas of messaging, security, and networking. Opus One has provided objective testing results for publication and private use since 1983.

This document is copyright © 2015 Opus One, Inc.

APPENDIX

DEFINITION OF TERMS

Spam is unsolicited commercial bulk email. We consider messages to be "spam" if there is no business or personal relationship between sender and receiver and which are obviously bulk in nature. Mail messages that may not have been solicited, but which show a clear business or personal relationship between sender and receiver, or are obviously a one-to-one message, even if unsolicited and unwanted, are not considered "spam."

Spam catch rate measures how well the spam filter catches spam. We have used the commonly accepted definition of specificity, which is the number of spam messages caught divided by the total number of spam messages received. The missed spam is one minus the spam catch rate.

False positive rate measures the number of legitimate emails misclassified as spam. Different vendors and testing services define false positive rate in different ways, typically either specificity or positive predictive value. In this report, false positive rate is defined using positive predictive value as (1 – ((messages marked as spam – false positives)/(total messages marked as spam))). The spam accuracy rate is one minus the false positive rate.

TESTING METHODOLOGY

Anti-spam products were evaluated by installing them in a production mail stream environment. The test simultaneously feeds the same production stream to each product, recording the verdict (typically "spam," "not spam," or "suspected spam") for later comparison.

Each product tested was acquired directly from the vendor or through normal distribution channels. Each product tested was under an active support contract, and was believed to be "up to date" with publicly released software and signature updates.

Where multiple versions were available from a vendor, the technical support team for each vendor was consulted to determine the "recommended" platform for use. To minimize confusion, products were not upgraded during the test cycle, although anti-spam and anti-spam engine updates were typically and automatically made by each product during the term of the test.

All systems were able to connect to the Internet for updates and DNS lookups. A firewall was placed between each product and the Internet to block inbound connections, while outbound connections were completely unrestricted on all ports.

Each product was configured based on the product manufacturer's recommended settings.

Where easily executed, multiple scenarios were used for a product, including a factory-default aggressive setting ("suspect spam"), and conservative setting ("certain spam"), based on the vendor's recommendation. In cases where obviously inappropriate settings were included by default, these settings were changed to support the production mail stream. "Maximum message size" -- to accommodate messages of varying sizes -- was the most commonly changed setting.

The tests drew on the real ".COM" corporate message stream because this message stream contains no artificial content and best represents the normal enterprise stream. No spurious spam or non-spam content was injected into the stream. No artificial methods to attract spam were employed.

Each product was connected to the Internet to retrieve signature and software updates as often as recommended by the vendor. If vendor technical support teams recommend a shorter update cycle, this recommendation was implemented.

Because products were not receiving email directly from the Internet, the reputation service of each product had to be individually configured to support the multi-hop configuration. In cases where products were unable to handle a multi-hop configuration with reputation service, the reputation service results were gathered at the edge of the network and then re-combined with the anti-spam results after the test was completed.

For many products, this re-combination better illustrates the actual performance a network manager would see and significantly changes the test results from a test which does not incorporate reputation service results.

Once the messages were received, Opus One manually read through every single message, classifying it as “spam,” “not spam,” or “unknown” according to the definitions above. All mailing lists which have legitimate subscriptions were considered “not spam,” irrespective of the content of any individual message. However, mailing list messages that were blocked for malware (for example, because they included only a URL to a malware site) were not considered false positives, even if this was just one message out of a digest of a mailing list.

Messages were classified as “unknown” if they could not be definitively categorized as “spam” or “not spam” based on content, or if they were so malformed that it could not be determined that they were spam, viruses, or corrupt software. All "unknown" messages were deleted from the data set, and do not factor into the result statistics. The total number of “unknown” messages in the sample was small, typically less than 0.1% of the total sample size.

Once the manual qualification of messages was completed, all results were placed in an SQL database. Queries were then run to create false positive and false negative (missed spam) lists. False positives and false negatives for each product were evaluated and any errors in the original manual classification were fixed. Once the data sets were determined to be within acceptable error rates, the databases were reloaded and the queries recreated.

Each anti-spam engine provides a verdict on messages. While this is often internally represented as a number, the verdict in most products is reduced to a categorization of each message as being “spam” or “not spam.” In many anti-spam products, a third category is included, typically called “suspected spam.”

In this test, products were configured at the factory-default settings, where possible, to have three verdicts (spam, suspect spam, and not spam). Where products have three verdicts, suspect spam is not considered to be spam. Thus, suspect spam was not included in the catch rate and false positive rate calculations. The one exception to this is Vendor D; in this product, “suspected spam” is defined differently and should be considered spam.

Catch rate refers to the number of spam messages caught out of the total number of spam messages received. When spam is not caught, it is called a false negative.

- False negative means the test said “this was not spam,” and it was.
- False positive means the test said “this was spam,” and it wasn’t.