Joel Snyder
Opus One
March 2011

COMPARING INDUSTRY-LEADING ANTI-SPAM SERVICES

RESULTS FROM 12 MONTHS OF TESTING

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.

All of the anti-spam solutions in Gartner's May 2010 "Leaders" and "Challengers" Magic Quadrant categories were tested. In total, eight vendors were evaluated over the course of a year. The only vendor mentioned by name is Cisco IronPort. The remaining vendor names have been obfuscated.

TEST METHODOLOGY

To ensure consistency and reliability, Opus One operated within the following parameters during the 12-month long analysis from January 2010 to December 2010:

- Approximately 10,000 messages were selected at random for testing each month, with a total of 125,018 messages in the final evaluation set
- 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
- Tested products were "up to date" with current released software and signature updates
- Messages were hand classified as "spam" and "not spam" to ensure data validity
- Messages counted on a per-recipient basis
- Each of the tested products included the vendor-recommended or integrated reputation service in the results

The test results reported here are taken from Opus One's continuing anti-spam testing program. With six 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.

TABLE OF CONTENTS

Introduction and Test Methodology.....1
Test Results.....2
Spam Catch Rate Results.....3
False Positive Results.....4
Summary.....4
Appendix.....5

TEST RESULTS

Cisco IronPort solution demonstrated the highest spam capture rate and the most accurate rate of detection. The results are remarkable given the tradeoff between spam capture and false positive rates. For example, a vendor can catch 100% of spam if they block every message but then the false positive rate would also be 100%, which is obviously unacceptable.

Cisco IronPort consistently outperformed the other vendors, with the highest spam capture rate in ten of the twelve months measured. In the other two months, only one other solution exhibited a higher spam capture rate than Cisco IronPort. However, the vendors' false positive rates were significantly higher those two months by an average 17 times than that of Cisco IronPort.

With missed spam 144% relative to the leader, Vendor F placed second, but generated a false positive rate three times that of the Cisco IronPort solution. Vendors C, E and G missed spam 168-200% relative to the leader, albeit with a false positives four to six times higher. Vendors D, A and B had higher accuracy but at the cost of a lower spam catch rate.

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

Comparative
Anti-Spam Efficacy

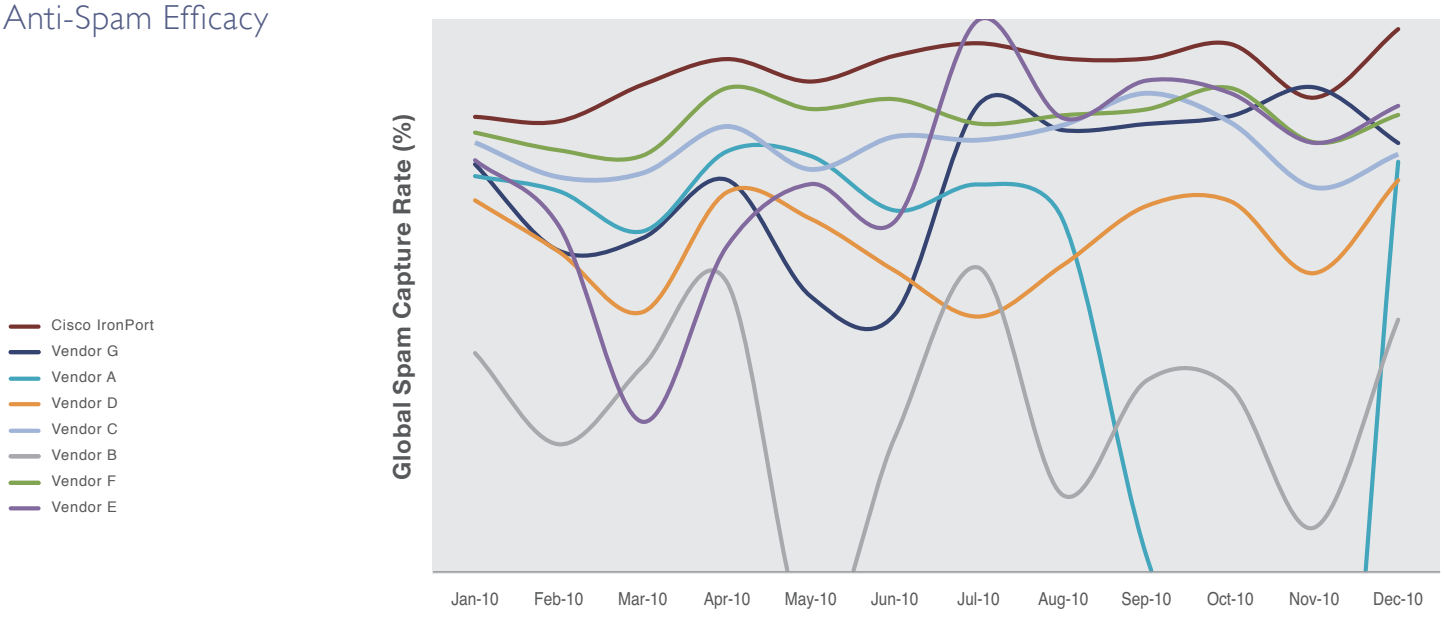
SPAM CATCH RATE RESULTS

The spam catch rate has a direct impact on end-users' satisfaction and productivity. With the high daily global volume of spam, even the slightest reduction in catch rates can have a major adverse effect. The relative catch rates for anti-spam vendors over the year-long period ending December 2010 are as follows:

Vendor	Missed Spam Relative to Leader
Cisco IronPort	**n/a**
Vendor F	**144%**
Vendor C	**168%**
Vendor E	**195%**
Vendor G	**199%**
Vendor D	**255%**
Vendor A	**344%**
Vendor B	**408%**

Month by month spam catch rate results by vendor over the testing period are graphed below. Because most anti-spam products have a high capture rate, the horizontal axis crosses at the 90% capture rate level.

Monthly Comparative Anti-Spam Efficacy



FALSE POSITIVE RESULTS

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. The relative results over the year-long period ending December 2010 are as follows:

Vendor	False Positives Relative to Leader
Cisco IronPort	**n/a**
Vendor B	**141%**
Vendor D	**172%**
Vendor F	**294%**
Vendor A	**320%**
Vendor G	**415%**
Vendor C	**452%**
Vendor E	**541%**

SUMMARY

Given the essential role of email in the operations of modern enterprises, spam poses a serious threat to their success. When a spam message finds its way into a user's inbox or a legitimate message is incorrectly identified as spam and quarantined, there is an immediate impact on productivity. While performance of the solutions evaluated in this analysis may vary by only a few percentage points, it's important to recognize that this difference can translate into hundreds, if not thousands, of unwanted and potentially problematic messages infiltrating a network.

Over the years, much ground has been gained in the battle against spam. Nevertheless, the number of threats continues to rise, demanding increasingly sophisticated and capable defense systems. The productivity of the global marketplace demands it.

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.

APPENDIX

DEFINITION OF TERMS

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." Testers defined as "spam" the messages for which there was no conceivable business or personal relationship between sender and receiver and which were obviously bulk in nature. Mail messages that may not have been solicited, but which showed a clear business or personal relationship between sender and receiver, or were obviously a one-to-one message, even if unsolicited and unwanted, were classified as "not spam." All mailing lists which have legitimate subscriptions were considered "not spam," irrespective of the content of any individual message.

Messages were classified as "unknown" if they were the result of virus double bounces, or 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 for each product were individually evaluated and any errors in the original manual classification were fixed. Because the number of false negatives is typically much higher (up to 700 false negatives per product are not unusual), the third party research firms did not evaluate every false negative for each product. Instead, testers evaluated each false negative for at least two different products, and sampled the false negative results for all other products to identify any errors in the original classification. 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 considered to be spam. As a result, suspect spam was included in the catch rate and false positive rate calculations.

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.