

# **Major Techniques for Classifying Spam**

By Paris Trudeau and Dr. Richard Cullen  
with Dave Zwieback



## Introduction

About 40 percent of the U.S. Postal Service mail is “business marketing”<sup>1</sup>, consisting mostly of unsolicited junk mail. That amounts to roughly 270 million pieces of mail each day<sup>2</sup>, or almost 100 billion per year. Annoying as junk mail is, the cost burden is mostly on the sender, since each piece of mail costs a discrete amount to produce and send.

Unlike postal mail, electronic mail is vastly less expensive for the sender. Even at bulk rates, it could cost a postal junk mailer \$250,000 to send out a million pieces; however a sender of unsolicited bulk e-mails—widely known as spam—can send millions of messages using a \$20-per-month dialup connection to the Internet.

The vast cost of spam is being borne by its recipients. By the end of this year, as much as 50 percent of all e-mail traffic is expected to be spam. An estimated 15 billion e-mail messages are currently sent worldwide per day<sup>3,4</sup>, so spam could soon account for more than 7 billion e-mail messages daily, or 2.6 trillion messages per year. AOL has recently corroborated the order of magnitude of this figure, stating that it has blocked 1 billion spam messages in just one day<sup>5</sup>.

To handle this enormous amount of data globally, a staggering amount of processing, bandwidth and storage capacity is required (2000 TB per day, which is equivalent to 1 million T1 circuits solely dedicated to delivering spam 24 hours a day). The human cost is equally stunning: assuming that it takes a person one second to delete a message, it will take humanity a combined 222 years to process just a day’s worth of spam.

The infrastructure costs along with the lost productivity caused by spam will add up to more than \$10 billion this year for U.S. organizations alone. In addition, spam may pose serious legal ramifications for its recipients, since, depending on the laws in the recipient’s locality, the contents of many spam messages are illegal<sup>6</sup>.

## Methods for Fighting Spam

From the very beginning of the spam epidemic, two complementary approaches of fighting it have emerged: technological and legal. While some states have already passed legislation<sup>7</sup>, there is still no US federal law on the books dealing with spam. In addition, since spam is truly a global phenomenon, with an increasing amount of messages crossing international borders in transit, international laws and treaties need to be implemented to effectively address the problem. In general, legal processes are lengthy, costly, and complicated, and do little to physically prevent someone from sending spam—or, more importantly, prevent someone from receiving it.

Fortunately, increasingly effective software tools that are capable of correctly identifying and filtering a large percentage of spam e-mails are becoming widely deployed. In this white paper, we will cover the most effective methods for capturing spam, on both transport and content levels. In addition to the transport-level filtering, we will specifically cover the following content-level methods of identifying spam:

- 1) Fingerprint Analysis
- 2) Lexical Analysis
- 3) Artificial Intelligence
- 4) Statistical Analysis
- 5) Heuristics

## Capturing spam at the transport level

Internet e-mail messages are transported between the senders and the recipients (or at least, their mail servers) via SMTP—the Simple Mail Transfer Protocol. Originally published in 1982, when the Internet was a fraction of its current size and the spam problem did not exist, this simple protocol allows anyone to use any SMTP mail server to send messages to anyone on the Internet, without authentication. As a result, e-mail messages are easily “spoofed”:

SMTP mail is inherently insecure in that it is feasible for even fairly casual users to negotiate directly with receiving and relaying SMTP servers and create messages that will trick a naive recipient into believing that they came from somewhere else.<sup>8</sup>

Although e-mail spoofing cannot be completely eradicated, it is possible to prevent spoofing for a particular domain by implementing user authentication. For instance, a mail server for company.com should deliver messages that appear to be from company.com users only after the users authenticate successfully (e.g., via ESMTP). However, this will not stop someone from forging an e-mail that appears to come from user@company.com and successfully delivering it to someone at hotmail.com.

Besides facilitating sending and receiving e-mail for its own users, an SMTP mail server could also act as an “open relay” for other users or mail servers, and attempt to deliver their mail as well. This suited the collaborative nature of the Internet 20 years ago, but these days the moment that an open relay is found by spammers, literally millions of spam messages could be sent through it. Thus, it is important to completely disable the open relay functionality of any Internet mail server—while the server’s internal users must be able to continue to send and receive messages to anyone they wish, external users (including spammers) must never be allowed to send messages through the mail server that are not addressed to that server’s users.

Once an open relay is discovered or setup by spammers, it is also usually found by (or can be reported to) the maintainers of one of the many public “blacklists”<sup>9</sup>

(also known as Realtime Blackhole Lists or RBLs) which contain updated and verified lists of IP addresses of such open relays. In addition, there are blacklists that contain domain names of known spammers. When receiving a mail connection, a server can check if it is coming from one of the blacklisted addresses or domains, and refuse it. Although such blacklists can reduce the amount of spam, they are also known to mistakenly list legitimate mail servers<sup>10</sup>. Thus, the use of publicly available blacklists should be complemented with the use of a whitelist appropriate to the particular organization's users. Alternatively, one can maintain a private blacklist based on the worst spam offenders for the organization's mail servers. In either case, black/white-listing can be an effective first defense in reducing the amount of spam.

## *Looking inside the messages*

### **Fingerprint Analysis**

The response rate for spam is only 0.01%<sup>11</sup> or less, so the spammers' business model forces them to send out millions of copies of the same e-mail, sometimes over a period of days or weeks. Theoretically, once someone receives a piece of spam, she can create a fingerprint of this message and share it with everyone. This way, anyone who receives this message fingerprint prior to receiving the message would be able to successfully identify and filter this message out.

Unfortunately, the messages often vary subtly for each recipient even within a single spam "campaign". For instance, spam e-mails commonly contain links that claim to allow the recipients to unsubscribe from future mailings, but in fact help the spammers identify and track the recipients and validate their e-mail addresses. In addition, creating fingerprints (also known as signatures or hashes) of e-mail messages could be a resource-intensive task, especially in high-volume enterprise environments. Fortunately, even after a casual analysis of spam, one can see that there are certain elements that would naturally be common—or even unique—to all "copy any DVD" campaign messages. Thus it is possible to extract and build a fingerprint database containing such unique elements.

This approach is similar to fingerprint-based identification of viruses. Like its virus-fighting counterpart, it is not always effective against rare or unusual spam messages. However, depending on the quality of the fingerprint database, and provided this database is kept current, fingerprint analysis methods are very effective in identifying common spam strains, and are also highly unlikely to incorrectly identify an innocuous message as spam.

## Lexical Analysis

While fingerprint analysis methods excel at preventing the spread of messages that are already known, like computer viruses, spam messages are constantly mutating to avoid detection. The perpetrators of spam are becoming increasingly technically savvy and aware of the anti-spam tools used against them. One of the ways in which they can avoid detection by fingerprint analysis tools is by using creative variations on the typical spam vocabulary. For instance, there might be a fingerprint for the phrase “get free Viagra” in the database, but not the almost nonsensical “V1agra for absolutely free can be obtained here”.

Nevertheless, any spam attempting to advertise free medicines is going to stand out from typical business communications, with the exception of, perhaps, messages within the pharmaceutical industry. However, even in such a context, legitimate messages with the names of medicines are unlikely to be anywhere near the word “free”, so lexical analysis can be successfully used to identify spam.

Lexical analysis works by examining the context for all of the words and phrases in a message. The presence of a particular suspicious word or phrase by itself does not necessarily mean that the message is spam. Instead, each word or phrase is assigned a weight depending primarily on the context in which they are found. For example, “Viagra” in the context of a discussion of its generic name “sildenafil citrate” would most likely be in a legitimate e-mail, while its presence anywhere near the word “free” is a good indicator that the message is spam. Once the whole message is analyzed, the weights for the found elements are combined, and the resulting score is compared to a preset threshold. If the score is above the threshold, the message is considered spam.

Lexical analysis can also be applied to catch variations of words and phrases—a technique which is becoming very popular among spammers. For instance, it is possible to catch not only “Viagra”, but also “V1agra” or “Vayagra”. The overall effectiveness and efficiency of lexical analysis algorithms in filtering spam is highly dependent on the quality of the rules, and their assigned weights. Naturally, while the lexical analysis algorithms are the same for any language, a separate rule and dictionary configuration is required for each language.

## Artificial Intelligence

Over time, every person exposed to spam learns to identify unwanted messages quickly and accurately, sometimes after reviewing only the subject line or the sender's e-mail address. Neural Networks is an Artificial Intelligence technique that simulates in software the way that a human brain learns to recognize various patterns. Neural Networks have been successfully implemented in character and speech recognition, intrusion and virus detection, and now in fighting spam.

A classic example of Neural Networks in action is a virtual mouse that “lives” in a maze with randomly placed chunks of cheese, some of which have gone bad. When the mouse finds a piece of cheese, either its hunger gets satisfied or it starts to feel sick, depending on whether it eats the fresh or the spoiled cheese. The mouse is initially trained to distinguish between the two types of cheese, and over time it learns to eat mostly the good cheese, eventually making very few errors.

In a basic configuration, a Neural Network has inputs, outputs, and the interconnections between them. As the name implies, inputs—or input nodes—represent the source data that needs to be analyzed. In the example of the virtual mouse, the input nodes could be the color, shape, or smell of a piece of cheese that would allow the mouse to distinguish the good cheese from the bad. For e-mail messages, the inputs could be all the words in a message.

The output nodes represent the results of the analysis. In each of the simplified examples above, there would be two output nodes: “fresh” and “spoiled” for the cheese, “non-spam” and “spam” in the case of e-mail. The number of outputs is not constrained to two—a Neural Network can be trained to organize e-mails into any number of categories (personal, business, spam, etc.).

Each input node is connected to each output node. This fully connected configuration represents the fact that all of the inputs—in the case of spam, all of the words found in an e-mail message—are being considered simultaneously. This is the key to the success of Neural Networks, since the overall context of the message is a better predictor of whether the message is spam than the individual words. As is the case with lexical analysis, if “Viagra” is found along with “sildenafil citrate”, the message is likely to be legitimate, whereas if the message contains “Viagra” and “free”, it is likely spam.

The accuracy of the Neural Network is extremely dependent on how well it is trained. In the process of training, a Neural Network is repeatedly fed both the input and output data, and the network adjusts the weights of the interconnections between the nodes to produce increasingly accurate results over time. In essence, the Neural Network is learning by example in a similar way that a child learns to distinguish between cats and dogs after a parent repeatedly points out the differences between them, and corrects any of the child's mistakes. Neural Networks can achieve a substantially high accuracy rate, provided a sufficiently large body of messages is supplied to the Neural Network during training. In addition, when spam messages slip through, or valid messages are incorrectly identified

as spam, a detection system based on Neural Networks can be easily trained to not make the same mistakes in the future <sup>12</sup>.

## Statistical Analysis

There is considerable overlap between the fields of Neural Networks and statistics <sup>13</sup>. Similarly to Neural Networks, running a large amount of correctly identified spam and non-spam messages through statistical analysis like Bayesian Filtering will produce a database of all the words found in the messages along with the probability that a particular word belongs to a spam message. For instance, a word like “promotion” might have a 99 percent probability of being in a spam message, while “obfuscation” might have a 99 percent probability of being in a non-spam message.

Once again, as with lexical analysis and Neural Networks, inspecting the whole message—the context—is much more meaningful and accurate than looking at individual words. Using Bayesian Filtering and the database of probabilities of the words generated during training, the overall probability that a particular message is spam can be easily computed, even if the database does not contain all of the words found in the message. Properly trained statistical filters can achieve accuracy rates similar to those of Neural Network-based filters. <sup>14</sup>

## Heuristics

Heuristics (from the Greek word “heuriskein”, to find) has been successfully used to identify unknown viruses, and is now widely used to fight spam. In the context of e-mail, heuristics is a method of applying successive tests to a message to determine if it is likely to be spam. Heuristics is not so much a unique method as a framework for combining various tests, and assigning relative scores to their results. The sum of the resulting scores of the tests performed on a message is compared to a preset threshold; if this threshold is reached, no further tests need to be performed. Tests that can be incorporated in a heuristic framework include all of the approaches described above (transport-level testing, as well as fingerprint, lexical, AI, and statistical methods). In addition to investigating contents of the e-mail, message attributes like time and date, size, number of attachments, MIME-types, etc., also can (and should) be tested. <sup>15</sup>

In an enterprise environment, the best way to structure tests within a heuristics framework is to perform the least resource-intensive and most accurate tests first, and only run successive tests if the previous ones return inconclusive or negative results (i.e., if the threshold is not reached). The larger the number of users (for instance, in a large enterprise), the more difficult it is to find consensus about what is considered spam, and thus the lower the potential accuracy of any one test. Furthermore, as is the case with viruses, because of the constant mutation of spam messages, it is imperative that two or more tests are performed on each message to ensure the highest rate of spam detection. In addition, an enterprise policy that clearly defines acceptable e-mail use, as well as centralized administration of e-mail filtering will further improve the accuracy of spam testing. The overall success rate of any heuristic framework clearly depends on the

accuracy of the tests that comprise it, as well as the effectiveness of the scoring and threshold mechanism.

## False Positives and Negatives

The definition of spam depends on an individual or organization, and it may change over time. Even if the definition of spam were always exact, no human or software can successfully identify spam 100 percent of the time. There will always be mistakes: false positives (legitimate e-mails marked as spam by the filtering software) and false negatives (spam e-mails that are not identified as such). A low percentage of false positive and/or negatives may be tolerable for some users, and completely unacceptable for others. A good spam filter should always be configured to not be too restrictive (as is the case with some publicly available RBLs), and at the same time not be too open. Furthermore, it is advisable to deliver suspicious messages with low probabilities of being spam to the users, albeit into a separate “suspected spam” folder. This way, the users have the ability to alert the administrators when specific messages are incorrectly identified, thus improving the long-term accuracy of spam detection. In general, the goal of any successful anti-spam configuration is to achieve the maximum accuracy while keeping the required human intervention to a minimum.

## Fighting Spam with the SurfControl E-mail Filter

SurfControl E-mail Filter is a highly effective, scalable enterprise solution for the spam epidemic. Once the E-mail Filter is running, it immediately protects the company mail server by disabling any open relay functionality, which can be abused by spammers to send millions of unsolicited and potentially illegal messages. Furthermore, the E-mail Filter can verify the integrity of the any e-mail message and ensure that the message has not been spoofed. In order to quickly reduce e-mail traffic from the worst offenders, the E-mail Filter supports multiple Real Time Blackhole Lists (RBLs), as well as custom black/white lists that can contain e-mail addresses, domains, or IP addresses.

At the heart of the E-mail Filter, is the Anti-Spam Agent (ASA). Based on the proprietary database containing more than 35,000 unique digital fingerprints, the ASA is able to positively identify millions of spam and junk e-mails, and other objectionable material not suited for the workplace. A dedicated content team is constantly updating the database, and each customer automatically receives new fingerprints at least once a day. Should a spam message get through, the user can forward it to the ASA team, which will update the fingerprint database.

For e-mails that are not yet registered with the ASA, the E-mail Filter’s Spam and Adult Dictionary and powerful LexiMatch lexical analysis engine can be used to detect e-mails that contain the most commonly used spam and adult words and phrases. Dictionaries in English, French, Dutch, Spanish, Italian, and German are included, and traditional and simplified Chinese and Japanese dictionaries are available. Administrators can easily customize the dictionaries, the LexiMatch rules, as well as use the advanced functionality

of the drag-and-drop Rules Administration interface to tailor the ruleset for the specific requirements of their organization.

Using scalable Neural Networks technology, the Virtual Learning Agent (VLA) provides unmatched accuracy in identifying adult spam messages out of the box, and can be easily trained to recognize a message of any type. Furthermore, using 22,000 Artificial Intelligence algorithms, the optional Virtual Image Agent (VIA) is able to intelligently analyze images found in e-mails and attachments, and filter out pornographic material.

Leveraging SurfControl's extensive content-filtering and anti-virus experience, the E-mail Filter implements the best available filtering technologies within an extensible and scalable heuristic framework, ensuring a spam-free enterprise.

## Conclusion

According to Ferris Research, by the end of this year the average e-mail user will receive more than 40 spam messages a day. With a thoughtful implementation of spam classification techniques, it is possible to filter out a large portion of the spam, and thus minimize the financial and legal exposure to the enterprise caused by the spam epidemic. In the long term, by reducing the amount of unsolicited bulk e-mails that are delivered to end-users, it is actually possible to drive a substantial part of the offending spammers out of business.

## Authors' Biographies

---

### Paris Trudeau

Paris is a Product Marketing Manager at SurfControl who has nearly ten years of experience in product promotions and marketing campaigns. In her capacity at SurfControl, she works in alliance with the engineering team to develop products and ensure that their features are customer-driven enhancements. Paris also works closely with SurfControl's marketing group throughout the development process to get the product properly positioned in the marketplace.

Paris received her B.A. in Economics at University of California - Irvine.

### Dr. Richard Cullen

Richard joined SurfControl in December 2000 during the acquisition of EmUTech P/L as a senior developer on the e-mail development team. In late 2001 he became Development Team Leader for the SurfControl e-mail filter products.

Richard brings with him several years experience in the software development industry, working in the UK, Switzerland and Australia for AutoDesk and anti-virus and security specialists Sophos.

Richard has a B.Sc. from the University of Birmingham and a Ph.D. from London University (Imperial College).

### Dave Zwieback

Dave is the Technical Director of inkcom ([www.inkcom.com](http://www.inkcom.com)), a company which specializes in Infrastructure (System, Network, Storage, and Security) Architecture. He can be reached at [zwieback@inkcom.com](mailto:zwieback@inkcom.com).

---

## Bibliography

- <sup>1</sup> Jonathan Krim, *Spam's Cost To Business Escalates*. WashingtonPost.com. March 13, 2003. <http://www.washingtonpost.com/ac2/wp-dyn/A17754-2003Mar12>.
- <sup>2</sup> United States Postal Service, *2002 Annual Report*. <http://www.usps.com/history/anrpt02/>.
- <sup>3</sup> Channel One, *Market Overview*. <http://www.one.ie/report/e-mail/marketoverview.asp>.
- <sup>4</sup> Cisco Systems, Inc., *State of the Internet*. <http://www.cisco.com/warp/public/779/govtaffs/factsNStats/stateinternet.html>.
- <sup>5</sup> BusinessWire, *Over One Billion Spam E-mails Now Blocked in One Day by AOL*, March 5, 2003. [http://biz.yahoo.com/bw/030305/55356\\_1.html](http://biz.yahoo.com/bw/030305/55356_1.html).
- <sup>6</sup> Scott Hazen Mueller, editor. *Why is spam bad?*, <http://spam.abuse.net/overview/spambad.shtml>.
- <sup>7</sup> Roy Mark, *Spam Law Foe Reverses Direction*, InternetNews.com, October 28, 2002. <http://dc.internet.com/news/article.php/1489211>.
- <sup>8</sup> J. Klensin, editor, *RFC 2821: Simple Mail Transfer Protocol*, April 2001. <http://www.ietf.org/rfc/rfc2821.txt?number=2821>.
- <sup>9</sup> *Google Web Directory*, <http://directory.google.com/Top/Computers/Internet/Abuse/Spam/Blacklists/>
- <sup>10</sup> Fred Langa, *Langa Letter: Real-Life Spam Solutions*, InformationWeek, Nov. 18, 2002. <http://www.informationweek.com/story/IWK20021115S0018>.
- <sup>11</sup> William S. Yerazunis, *Sparse Binary Polynomial Hashing and the CRM114 Discriminator*, Spam Conference Presentation, January 2003. [http://crm114.sourceforge.net/crm\\_slides/img39.html](http://crm114.sourceforge.net/crm_slides/img39.html).
- <sup>12</sup> Z Solutions, Inc., *Light Description of Neural Networks*, <http://www.zsolutions.com/light.htm>.
- <sup>13</sup> Warren S. Sarle, editor, *Neural Network FAQ*, 2002. [ftp://ftp.sas.com/pub/neural/FAQ.html#A\\_stat](ftp://ftp.sas.com/pub/neural/FAQ.html#A_stat).
- <sup>14</sup> Paul Graham, *A Plan for Spam*, August 2002. <http://paulgraham.com/spam.html>.
- <sup>15</sup> Stephen M. Sladaritz Sr., *About Heuristics*, SANS Info Sec Reading Room, March 23, 2002. <http://www.sans.org/rr/malicious/heuristics.php>.