

## How Spammers Fool Spam Filters

by: Dr. Paul Judge, 07/06/2005

<http://www.securitydocs.com/library/3436>

Effectively stopping spam over the long-term requires much more than blocking individual IP addresses and creating rules based on keywords that spammers typically use. The increasing sophistication of tools spammers use coupled with the increasing number of spammers in the wild has created a hyper-evolution in the variety and volume of spam. The old ways of blocking the bad guys just don't work anymore.

Examining spam and spam-blocking technology can illuminate how this evolution is taking place and what can be done to combat spam and reclaim e-mail as the efficient, effective communication tool it was intended to be.

There are several widely-used methods for filtering spam, each of which can be defeated by spammers to some degree. Understanding the strengths and weaknesses of each approach and the methods spammers use to defeat them is the basis of an effective, comprehensive anti-spam strategy.

### Signature-based Filters

Signature-based filters examine the contents of known spam, usually derived from honey pots, or dummy e-mail addresses set up specifically to collect spam. Once a honey pot receives a spam message, the content is examined and given a unique identifier. The unique identifier is obtained by assigning a value to each character in the e-mail. Once all characters have been assigned a value, the values are totaled, creating the spam's signature. The signature is added to a signature database and sent as a regular update to the e-mail service's subscribers. The signature is compared to every e-mail coming in to the network and all matching messages are discarded as spam.

The benefit of signature-based filters is that they rarely produce false-positives, or legitimate e-mail incorrectly identified as spam.

The drawback of signature-based filters is that they are very easy to defeat. Because they are backward-looking, they only deal with spam that has already been sent. By the time the honey pot receives a spam message, the system assigns a signature, and the update is sent and installed on the subscribers' network, the spammer has already sent millions of e-mails. A slight modification of the e-mail message will render the existing signature useless.

Furthermore, spammers can easily evade signature-based filters by using special e-mail software that adds random strings of content to the subject line and body of the e-mail. Because the variable content alters the signature of each e-mail sent by the spammer, signature-based spam filters are unable to match the e-mail to known pieces of spam.

Developers of signature-based spam filters have learned to identify the tell-tale signs of automated random character generation. But as is often the case, spammers remain a step ahead and have developed more sophisticated methods for inserting random content. As a result, most spam continues to fool signature-based filters.

### Rule-based (Heuristic) Filtering

Rule-based filters scan e-mail content for predetermined words or phrases that may indicate a message is spam. For example, if an e-mail administrator includes the word "sex" on a company's rule-based list, any e-mail containing this word will be filtered.

The major drawback of this approach is the difficulty in identifying keywords that are consistently indicative of spam. While spammers may frequently use the words "sex" and "Viagra" in spam e-mails, these words are also used in legitimate business

correspondence, particularly in the healthcare industry. Additionally, spammers have learned to obfuscate suspect words by using spellings such as "S\*E\*X", or "VI a a GRR A".

It is impossible to develop dictionaries that identify every possible misspelling of "spammy" keywords. Additionally, because filtering for certain keywords produces large numbers of false positives, many organizations have found they cannot afford to rely solely on rule-based filters to identify spam.

## Blacklists

The goal of blacklisting is to force Internet Service Providers (ISPs) to crack-down on customers who send spam. A blacklisted ISP is blocked from sending e-mail to organizations. When an ISP is blacklisted, they are provided with a list of actions they must take in order to be removed from the blacklist. This controversial method blocks not just the spammers, but all of the ISP's customers. Blacklisting is generally considered an unfriendly approach to stopping spam because the users most affected by the blacklist are e-mail users who do not send spam. Many argue blacklisting actually damages the utility of e-mail more than it helps stop spam since the potential for blocking legitimate e-mail is so high.

In addition to the ethical considerations, there are other problems with blacklists. Many blacklists are not updated frequently enough to maintain effectiveness. Some blacklist administrators are irresponsible in that they immediately block suspect servers without thoroughly investigating complaints or giving the ISP time to respond. Another downside is that blacklists are not accurate enough to catch all spam. Only about half of servers used by spammers, regardless of how diligent the blacklist administrator may be, are ever cataloged in a given blacklist.

Blacklists are used because they can be partially effective against spammers who repeatedly use the same ISP or e-mail account to send spam. However, because spammers often change ISPs, re-route e-mail and hijack legitimate servers, the spammer is a moving target. Blacklist administrators are forced to constantly revise lists, and the lag-time between when a spammer begins using a given server and when the blacklist administrator is able to identify the new spam source and add it to the blacklist allows spammers to send hundreds of millions of e-mails. Spammers consider this constant state of flux a part of doing business and are constantly looking for new servers to send spam messages.

Blacklists, therefore, have some utility in stopping known spammers. Because of their limitations, however, this data should only be used in conjunction with other sources to determine if a given message is spam.

## Whitelists

Whitelists are databases of trusted e-mail sources. The list may contain specific e-mail addresses, IP addresses or trusted domains. E-mails received from a whitelisted source are allowed to pass through the system to the user's email box. The list is built when users and e-mail administrators manually add trusted sources to the whitelist. Once built, the catch-rate for spam can be close to 100%, however, whitelists produce an inordinate number of false positives.

It is virtually impossible to produce an exhaustive list of all possible legitimate e-mail senders because legitimate e-mail can come from any number of sources. To get around this difficulty, some organizations have instituted a challenge-response methodology. When an unknown sender sends an e-mail to a user's account, the system automatically sends a challenge back to the sender. Some challenge-response systems require the sender to read and decipher an image containing letters and numbers. The image is designed to be unreadable by a machine, but easily recognizable by a human. Spammers would not spend the time required to go through a large number of challenge-response e-mails, so they drop the address and move on to those users who don't use such a system.

Whitelists are only partially successful and impractical for many users. For example, problems can arise when users register for online newsletters, order products online or register for online services. If the user does not remember to add the new e-mail source to their whitelist, or if the domain or source is entered incorrectly, the communication will fail. Additionally, whitelists impose barriers to legitimate e-mail communication and are viewed by some as just plain rude.

Whitelists are not widely used by e-mail users and administrators as a primary tool to fight spam because of the high number of false positives, and the difficulties in creating a comprehensive list of e-mail sources. Because whitelists are not widely used, spammers typically do not develop countermeasures. As with other spam fighting techniques, whitelists are most effective when used in conjunction with other anti-spam tools.

## Bayesian Filters

Named after Thomas Bayes, an English mathematician, Bayesian Logic is used in decision making and inferential statistics. Bayesian Filters maintain a database of known spam and ham, or legitimate e-mail. Once the database is large enough, the system ranks the words according to the probability they will appear in a spam message.

Words more likely to appear in spam are given a high score (between 51 and 100), and words likely to appear in legitimate e-mail are given a low score (between 1 and 50). For example, the words "free" and "sex" generally have values between 95 and 98, whereas the words "emphasis" or "disadvantage" may have a score between 1 and 4.

Commonly used words such as "the" and "that", and words new to the Bayesian filters are given a neutral score between 40 and 50 and would not be used in the system's algorithm.

When the system receives an e-mail, it breaks the message down into tokens, or words with values assigned to them. The system utilizes the tokens with scores on the high and low end of the range and develops a score for the e-mail as a whole. If the e-mail has more spam tokens than ham tokens, the e-mail will have a high spam score. The e-mail administrator determines a threshold score the system uses to allow e-mail to pass through to users.

Bayesian filters are effective at filtering spam and minimizing false positives. Because they adapt and learn based on user feedback, Bayesian Filters produce better results as they are used within an organization over time.

Bayesian filters are not, however, foolproof. Spammers have learned which words Bayesian Filters consider spammy and have developed ways to insert non-spammy words into e-mails to lower the message's overall spam score. By adding in paragraphs of text from novels or news stories, spammers can dilute the effects of high-ranking words. Text insertion has also caused normally legitimate words that are found in novels or news stories to have an inflated spam score. This may potentially render Bayesian filters less effective over time.

Another approach spammers use to fool Bayesian filters is to create less spammy e-mails. For example, a spammer may send an e-mail containing only the phrase, "Here's the link...". This approach can neutralize the spam score and entice users to click on a link to a Web site containing the spammer's message. To block this type of spam, the filter would have to be designed to follow the link and scan the content of the Web site users are asked to visit. This type of filtering is not currently employed by Bayesian filters because it would be prohibitively expensive in terms of server resources and could potentially be used as a method of launching denial of service attacks against commercial servers.

As with all single-method spam filtering methodologies, Bayesian filters are effective against certain techniques spammers use to fool spam filters, but are not a magic bullet to solving the spam problem. Bayesian filters are most effective when combined with other methods of spam detection.

## The Solution

When used alone, each anti-spam technique has been systematically overcome by spammers. Grandiose plans to rid the world of spam, such as charging a penny for each e-mail received or forcing servers to solve mathematical problems before delivering e-mail, have been proposed with few results. These schemes are not realistic and would require a large percentage of the population to adopt the same spam eradication method in order to be effective.

Working alone, each individual spam-blocking technique works with varying degrees of effectiveness and is susceptible to a certain number of false positives. Fortunately, the solution is already at hand. A secure e-mail gateway appliance can provide a highly accurate solution by correlating the results of single-detection techniques with a correlation engine thereby offering complete protection.

## About the Author

Dr. Paul Judge is a noted scholar and entrepreneur. He is Chief Technology Officer at CipherTrust, the industry's largest provider of enterprise email security. The company's flagship product, IronMail provides a best of breed defense against [phishing attacks](#) and other email-based threats. Learn more by visiting [www.ciphertrust.com](http://www.ciphertrust.com) today.