

# How URL Spam Filtering Beats Bayesian/Heuristics Hands Down

By Ted Green, President, Greenview Data, Inc.

## Abstract

The evolution and sophistication of spam and spammers has sparked much debate as to which spam filtering methods are the most accurate. This whitepaper addresses issues and problems associated with Bayesian and Heuristic filtering solutions and argues that URL ("click me" link) filtering is the most accurate and predictable method of blocking spam. URL filtering does not require customer tuning and eliminates the need for Bayesian/Heuristic solutions which are error prone and unpredictable. Implementing URL filtering with conservative IP blacklists and very limited and specific heuristic filtering has the ability to solve spam with over 99% blocking, miniscule false-positive and without any customer tuning.

# Table of Contents

<b>Abstract .....</b>	<b>1</b>
<b>Introduction .....</b>	<b>3</b>
<b>Traditional Anti-Spam Techniques .....</b>	<b>4</b>
IP Blacklists .....	4
Content Analysis (Heuristics / Bayesian) .....	5
Effectiveness of Traditional Anti-spam Techniques .....	6
<b>The Advantages and Problems of URL Filtering .....</b>	<b>6</b>
<b>How SpamStopsHere solves the URL Filtering problems .....</b>	<b>7</b>
<b>The Predictability of URL Filtering .....</b>	<b>8</b>
<b>The Unpredictability of Bayesian / Heuristic Filtering .....</b>	<b>9</b>
<b>Phrase Filtering Augments URL Filtering .....</b>	<b>9</b>
<b>Using URL Filtering to Solve Spam .....</b>	<b>9</b>

## Introduction

Spam has evolved from a mere annoyance to a serious IT threat. It is widely recognized that spam can cause thousands of dollars per employee in lost productivity. Allowing pornographic spam to enter the corporate network now poses EEOC liability risks. Phishing scams can result in identity theft and devastating consequences for employees. Spam, viruses and spyware are becoming ever more intertwined.

Today's anti-spam solutions primarily rely on two general techniques:

- IP-Blacklists to block spam at its source
- Content Analysis (Heuristics) to block spam based on spam characteristics and likelihood from an examination of the headers and body

These techniques achieve reasonable, but not exceptional results, with the best such systems, often using both techniques, blocking 95% of spam. A primary concern with any anti-spam system is its likelihood of blocking a legitimate email. If a system blocks even 1 in 1000 legitimate emails (0.1% false positive rate) users are concerned that an important email might be blocked and therefore review their blocked emails. We argue that any system that requires review of the blocked spam is no solution at all - users are looking at spam. While aggressive, highly tuned Heuristics systems might block 99% of spam, they then block 1% of legitimate emails, causing users to lose confidence. Even used together, these techniques reduce spam, but don't quite solve it.

In this whitepaper we review traditional spam filtering techniques, and then detail the advantages of a long sought anti-spam technique called "URL filtering". Instead of blocking spam at its source, it blocks spam based on its intended destination, the "Click me" links found in nearly all spam messages. Spammers want recipients to act on their message in some way - usually by clicking on a link to visit a website or calling a phone number.

While almost "obvious" as a spam solution, previous attempts to implement URL filtering have had limited success, due to the elusiveness of spammers. We will explain the problems in implementing URL filtering and how SpamStopsHere solved them to create an exceptionally accurate commercial system which for two years running has blocked over 99% of spam that contains a URL. SpamStopsHere uses a large database of these "Click me" links and phone numbers to identify spam with absolute certainty and miniscule risk of false-positives.

Most, but not all spam contains a blockable URL; approximately 5 - 10% does not, such as spam advertising penny stocks. More traditional methods are still needed to block non-URL spam. However, by knowing that the URL filtering level has already blocked over 90% of all spam, very conservative methods can be used to ensure a miniscule false-positive rate.

With its highly effective URL filtering, SpamStopsHere requires absolutely no customer tuning to achieve its exceptional accuracy. With a typical false-positive rate of 0.001% (1 in 100,000) few customers feel the need to review the blocked emails.

This whitepaper argues that URL filtering and other database-driven solutions are the future of anti-spam. Heuristic filtering has a minor role and Bayesian training of heuristic systems is error prone and unnecessary. URL filtering is very analogous to common anti-virus technology - a database driven system with nearly zero false-positives. No one needs to "train" their anti-virus system and this is now also true for anti-spam from SpamStopsHere.

We acknowledge that URL filtering by itself is not a total solution. Besides the 5% of spam without a blockable URL, URL filtering is not effective against email harvesting which has little or no content. Also, high volume email services may consider URL filtering too CPU intensive, requiring additional servers.

Combining URL filtering with conservative IP blacklists and very limited and specific heuristic filtering has the ability to solve spam with over 99% filtering, miniscule false-positive and without customer tuning.

- IP Blacklist - to block absolutely known spam sources, email harvesting and handle the needs of high volume email services
- URL Filtering - block the majority of spam, update the IP blacklists
- Heuristic rules - block the remaining non-URL spam, update the IP blacklists

This combined system is supplied by the anti-spam vendor and requires no customer tuning.

## Traditional Anti-Spam Techniques

Most anti-spam systems try to block spam through two techniques: 1) blacklists of known spam sources and/or 2) content analysis (heuristics) to identify typical spam words or phrases.

### IP Blacklists

Many traditional anti-spam systems use "blacklists" to block messages from previously identified spam sources. Blacklists are usually fully automated systems, called Real-time Blacklists, that are triggered when a spam is sent to any one of thousands of special email harvesting addresses. At that point all emails from that source will be blocked for one day, perhaps much longer. These blacklists run with little human review. Some blacklists rely on customer feedback (group voting) to identify spam. However, this is very error prone because many users will report "unexpected" email as spam; that which is one person's favorite monthly newsletter is another person's spam.

Spammers partially defeat blacklists by sending a single spam campaign from thousands of sources around the world; often from computers they have "hacked" using viruses or other tools. Thousands of small companies in third world countries are also in the business of sending spam. The current Internet defines about 2 billion possible sources from which a spam (or other email) could come. A good blacklist might be blocking 100 million of these sources. However, since over 100,000 additions and deletions need to be made every day, and additions are not made until after a new spam campaign starts, blacklists are error-prone, typically blocking only 80-90% of spam.

Besides missing some spam, automated real-time blacklists have an intrinsic flaw which guarantees that they will block some legitimate emails (false-positives). Consider the case of a small ISP with local business customers. One customer decides to promote his business by purchasing an email list and mailing out perhaps 100,000 emails. Very likely one of those is a special "harvesting" email address which will automatically trigger a real-time blacklist. The result is that all emails from that ISP will now be blocked by anyone using that blacklist; the email from all of the ISP's good customers will be blocked. The problem is compounded by the thousands of "free email" services available; many of these are overseas and make money by sending spam on the side.

Another type of blacklist, such as that originally created by Mailabuse.org (MAPS), and recently acquired by Trend Micro, uses a carefully researched list of IP addresses which excludes legitimate ISPs, corporations and other trusted sources. While such blacklists claim an extremely low false-positive rate, they have no ability to block even the most obvious spam sent from trusted sources. For this reason, spammers have recently concentrated on compromising "trusted" sources which are not on these lists.

The main problem with blacklists is that they have no ability to differentiate an obvious spam from an obvious legitimate email. For example, if someone is using the ISP above and sends a simple email of "Bob: Please pick up Linda from soccer practice", it may well be blocked for no apparent reason. Alternatively, if an obvious vulgar spam is sent from an unblocked (or trusted) source, it will not be blocked by the blacklist. The "obvious" errors made by blacklists give users low confidence in such systems.

## **Content Analysis (Heuristics / Bayesian)**

A second anti-spam technique employs "heuristics" to examine the email message for spam characteristics. It might look for hundreds of characteristics and each one is assigned a "weight"; if the sum of the weights exceeds a threshold, it is considered spam. While some characteristics relate to very technical details within the email header, many relate to the email message itself. Words like "Viagra", phrases like "free offer" and profanity will have a very high weight, perhaps high enough by itself to block the email. Other characteristics typically include large letters, red letters, many periods between letters (as in "F.r.e.e o.f.f.e.r"), existence of an image, etc.; hundreds in all.

Heuristic systems are not foolproof, particularly because of the techniques spammers use to defeat them. For example, spammers will obfuscate trigger words with deliberate misspellings, by adding invisible letters to the message, and with new tricks every week. This requires heuristic systems to be updated regularly (at least weekly) with new "rules" and weights.

Perhaps the most widely used heuristic system is the "free" software SpamAssassin. Here is how reviewer Logan G. Harbaugh described it for InfoWorld magazine in July 2004:

"Once I got the SpamAssassin software configured and running, its default settings provided acceptable performance, blocking 88 percent of spam, but with a very high 14.77 percent false-positive rate. With a few months of use and tuning, however, I expect its performance would improve substantially. Adding available plug-ins, such as the Bayesian filter or the content-checking filter, would likely help too."

(Considering that SpamStopsHere was reviewed by Network Computing Magazine at that time as blocking over 95% with less than 0.5% false-positive rate, it may seem difficult to believe that this reviewer would say "acceptable performance".)

When first installed, heuristic systems typically do not perform even reasonably well until they have been tuned for their company and even individual users. This tuning is often a very time-consuming process; over a period of a year, a medium sized company may spend over a thousand hours of staff time fine-tuning the anti-spam system. In some ways they are replacing the spam problem with an anti-spam problem.

Better heuristic systems use something called "Bayesian" algorithms to help automate the tuning of anti-spam systems. Based on user feedback of missed spams and false-positives, the system will automatically change its "weights" in hopes of becoming more accurate in the future. However the result of all this fine-tuning is difficult to predict; it often starts blocking legitimate

emails for no apparent reason or misses "obvious" spams.

To save staff time, some companies simply leave the heuristic anti-spam system at its factory default settings, which may be updated on a daily or weekly basis. However, serious spammers often test their spam campaigns against popular heuristic systems in order to defeat them.

The basic premise of heuristic anti-spam systems is that Artificial Intelligence can defeat spam. However, in reality, humans are still smarter than computers and one should never underestimate the intelligence and determination of spammers.

## **Effectiveness of Traditional Anti-spam Techniques**

While the best blacklists will block 90-95% of spam, they will miss some very obvious and vulgar spam, upsetting the recipients.

While heuristic systems are more likely to block obvious and vulgar spam, their complex ever-changing rules and weights leads to unpredictable results and false-positives.

Not surprisingly, heuristics system seem to have plateaued - the best tuned systems will block 95% of spam with a 0.01% false-positive rate, or block 99% of spam with a 1% false-positive rate. And these are company-wide averages; some users may experience only 70% spam blocking or a 10% false-positive rate because the system is tuned for the average user.

Perhaps the most frustrating aspect of heuristic systems is their unpredictability - if the weights are changed or new rules are added to block a new spam campaign, no one can be sure that it won't suddenly cause false-positives, and no one can be sure how it will affect different users.

Heuristic systems are unsuitable for some industries like medicine and law. "Viagra" and anatomical references are legitimate terms in medical correspondence. Similarly, law firms cannot risk blocking evidentiary materials that contain vulgar or offensive language.

With 95% spam blocking, many users will still receive several spam per day. Even a 1% false positive rate represents a major problem. If there is a one in 100 chance that a legitimate email is blocked, staff will spend time reviewing blocked spam to retrieve those few, but possibly important emails that were improperly blocked. This is a laborious, ongoing task that can assume greater demands than just manually deleting spam in the first place.

Anti-spam systems that require customers to constantly fine-tune the system waste valuable staff time, negating their supposed savings. Systems with a false-positive rate that causes customers to double-check the blocked emails are no solution at all - the spam is still being seen and then deleted.

## **The Advantages and Problems of URL Filtering**

While a single spam campaign might be sent from thousands of locations around the world in order to defeat IP-blacklists, it is much more difficult for a spammer to disguise the "click me" destination website. The website requires a unique domain name (URL) and each domain name costs money; each website also requires a spammer-friendly server on which to run it, and requires some effort to set up. Therefore, most spam campaigns have a "click me" link to a single website, or at most a dozen websites.

Consider that a single spam campaign might be sent from ten thousand locations, that each of the millions of spams might have slightly different content, but that each spam has the same "click me" link. Therefore, the SpamStopsHere designers decided to concentrate on identifying

the "click me" links of spammers and, as much as possible, ignore the source of the spam and the rest of the content.

Anti-spam companies have long sought to block spam through their "click me" destination website addresses. This has been the "holy grail" technique because it is so foolproof and almost completely eliminates the problems of false positives. However it has not been easy to achieve:

- Spammers constantly register new domain names and immediately begin using them for spam. (A single spammer might register hundreds of new domain names each week.)
- While there are "only" 2 billion possible spam sources, there are an infinite number of possible destination domain names.
- While automated systems can block source IP addresses based on various criteria, including customer spam reports, automated systems cannot easily determine which domain names to block.
- For example, if a particular spam is reported, an automated system can automatically add the IP address of the source to its blacklist.
- However, a spam may contain both legitimate domain names and the spammer's domain name. There is no way for an automated system to determine which to block and which not to block. Incorrectly blocking a legitimate domain, such as [www.ibm.com](http://www.ibm.com) would cause a disastrous number of false-positives.
- Attempting to use customer feedback (group voting) to identify spam is extremely error prone as many users will report "unexpected" email as spam. In the process, legitimate newsletters from many established companies will be blocked.

Largely due to the infinite number of possible "click me" domain names, the addition of thousands of new spammer domains every day, and the inability for automated systems to block the spammer domain names, most anti-spam companies consider URL filtering to be "theoretically best", but impossible to fully implement.

## How SpamStopsHere solves the URL Filtering problems

SpamStopsHere has perfected URL filtering to the point that it can now block nearly 99.9% of spam which contains a URL ("click me" link). This has been accomplished by developing several break-through, non-obvious solutions to the problems of URL filtering, and by testing and fine-tuning them over a period of two years against billions of emails. The solutions include:

- A highly trained 24/7 spam review staff that manually adds the majority of the URLs and verifies the automatically added URLs.
- Email harvesting of a 10-year old domain ([vedit.com](http://vedit.com)) that receives over one million spam per day. Combined with other email harvesting accounts (honey pots), nearly every new spam campaign is harvested.
- Patent pending technology which tracks the domain registration habits of thousands of known spammers. By instantly receiving domain registration information of new URLs, SpamStopsHere can block most spam campaigns before they even start, with zero false-positives.

- Patent pending technology that automatically detects new spam campaigns (which were not already anticipated) as they hit our filter servers. By comparing each email's URLs against the information in our databases, the system will either:
    - Automatically determine the spammer's URL, block it in real-time and have our 24/7 staff confirm it within a few minutes.
    - Pass the email as legitimate, but treat it as suspicious and send a copy to our 24/7 staff for human review and analysis. If it is spam, the staff will determine the best way to block it. The spam campaign will then be blocked within a few minutes.
- Note:** To protect the confidentiality of our customer's email, our staff only reviews emails sent to the email harvesting accounts which we own, and from select customers that have given us express permission to monitor for suspicious spam. Our service is fully HIPAA compliant. Large databases of known legitimate domains reduce the possibility that a link to an established company is ever blocked.
- All filter servers are updated every five minutes with the latest URLs needed to block all spam campaigns.

Two important elements of the solution are the 24/7 staff and the need to update the URL databases every five minutes.

As explained earlier, it is impossible for a fully automated system to determine which URLs to block. However, by combining patent pending semi-automated systems with highly trained staff, SpamStopsHere has perfected URL filtering.

Due to the thousands of new URLs used by spammers every day, it is necessary to update all filtering servers every five minutes. SpamStopsHere has developed the scalable technology to update hundreds of servers, verify that they are updated and fully monitor their performance.

## The Predictability of URL Filtering

While changes (fine-tuning) to a heuristic system often have unpredictable consequences, additions to URL filtering are absolutely predictable - it will block one spam campaign and nothing else.

For example, consider a legitimate newsletter from drugstore.com (a legitimate retailer) that advertises various health products and perhaps has "free" offers. Many heuristic systems will have trouble accepting this as a legitimate email due to "spam-like" content. Because SpamStopsHere almost completely ignores normal content, this email would not be blocked.

Now consider a spammer that takes the drugstore.com newsletter and changes all URL links from drugstore.com to drugstorerx.com (assuming this is the spammer's domain and website), and then sends this to a huge email list. This would be a heuristic system's nightmare. First the spammer's newsletter would likely not be blocked; then after many user reported the spam, the legitimate newsletter would also be blocked in the future.

With URL filtering, only the drugstorerx.com domain needs to be added to the blocking database. If not already in the blocking database, the SpamStopsHere technology would likely add it automatically and then have its 24/7 staff confirm it.

With URL filtering, the legitimate drugstore.com newsletter will never be blocked while the spammer's newsletter (with nearly identical content) will be blocked 100%.

Also, with URL filtering, the anti-spam vendor can determine precisely what will be blocked by policy. For example, the vendor can decide to block all emails that link to pornographic, casino and betting sites. Without blocking even vulgar personal emails, or discussions about casinos.

## **The Unpredictability of Bayesian / Heuristic Filtering**

We already explained how a Heuristic system might have trouble differentiating between a legitimate newsletter and a spammer's near identical version of it. Bayesian systems that automatically tune based on customer feedback of "spams" and "not spams" are even more unpredictable and error-prone. Part of the problem is not technology, but rather that normal users are "voting" on what is spam. Frankly, many users report legitimate newsletters as spam simply because they are not expecting it.

In the experience of SpamStopsHere, approximately one-half of the emails reported by customers as "spam" are not spam. Airline mileage reports are routinely reported as spam, follow-up emails from well know companies, and much more are reported as spam. Since these spam reports are rejected by our trained staff, they do not affect our accuracy. Bayesian systems that accept these inaccurate spam reports without trained human review will be highly skewed.

While even the best anti-spam system will occasionally miss the obvious spam, we will argue that a system using URL filtering and conservative IP blacklists is more accurate in spam filtering than the average user.

Therefore, trained staff at the anti-spam vendor must analyze all spam reports and determine what is spam and the best means of blocking it.

## **Phrase Filtering Augments URL Filtering**

While URL filtering can block the majority of spam, not all spam contains a URL. For example, penny-stock spams only mention the stock's symbol, followed by (unlikely) claims of how high it will rise. There are no blockable "click me" links. Nigerian scams only have an email reply address and occasionally a phone number. University "diploma" spams often consist of only an image which contains a phone number. None of these can be blocked by URL filtering.

URL filtering is well augmented by "Phrase Filtering", which as its name implies, simply checks the email for know spam phrases. It does not block based on single words, but rather on entire sentences that are distinctive in specific spam campaigns. Typical phrases might be "Diplomas from non-accredited universities", or "Watch abcd:pk explode on Monday". Again, trained staff can determine the best phrase to block the spam. In the experience of SpamStopsHere, only about 5000 phrases are needed at one time and most phrases are deleted after about one month.

## **Using URL Filtering to Solve Spam**

As mentioned before, URL filtering by itself will not completely solve the spam problem because not all spam contains a URL. It is estimated that 95% of spam contains a URL, leaving 5% which

must be blocked in other ways. While conventional techniques are needed to block the remaining 5%, very conservative phrase filtering and IP blacklists can be implemented since 95% of spam is already blocked by the URL filtering. It has been shown that IP blacklists heuristics which are set conservatively to only block 80% of spam can have miniscule false-positives. Combined with the URL filtering, over 99% of all spam can be blocked, without false positives, and without customer tuning.

Email harvesting attacks often contain nearly no content and therefore cannot be blocked by URL filtering, distinctive phrases or body content. Therefore, we argue that the following combination of methods will solve spam once and for all:

- URL Filtering - to block all spam with an "action", a "Click me" link or phone number.
- IP source reputation filtering to block spam from known sources (overseas spam servers). This is especially effective at blocking email harvesting attacks, the main weakness of URL filtering.
- Phrase filtering - to block non-URL spams based on distinctive phrases. (Using fairly straight forward pre-processing, attempts to obfuscate the phrases can be handled.)
- Image database filtering - to block image-only (non-clickable) spams.

It should be noted that this is an entirely database-driven solution. If the "signature" (e.g. IP address of sender, URL or phrase) is in the database, the email is blocked; otherwise it is not blocked. Unlike heuristic systems, there is no "grey area". The databases are maintained by the anti-spam vendor and all items in them are added or confirmed by trained staff.

URLs identify spammers with certainty and the heart of the eventual anti-spam solution is URL filtering. If spammers cannot get "action" emails to recipients, the financial incentive to send spam will be eliminated.

## **About Greenview Data:**

Greenview Data, Inc. has been providing its critically acclaimed SpamStopsHere network security solutions to clients across the globe since 2002. GDI also created and developed the powerful VEDIT tm text editor, which has been licensed to over 150,000 users since 1980. EBCDIC to ASCII conversion makes up the third division of Greenview Data; providing conversion solutions through turn-key contracting and consulting. Through the growth of its SpamStopsHere hosted service, and exciting R&D projects, Greenview Data, Inc. is looking to the future and another successful 25 years. GreenView Data is headquartered in Ann Arbor, Michigan. For more information, call (800) 458-3348 or visit [www.SpamStopsHere.com](http://www.SpamStopsHere.com)