Spam Filtering

A WORD TO THE WISE WHITE PAPER | BY LAURA ATKINS, CO-FOUNDER

Introduction

Spam filtering is a catch-all term that describes the steps that happen to an email between a sender and a receiver to distinguish between wanted and unwanted messages. Filtering companies and ISPs consider the underlying mechanisms behind spamfilters confidential, so senders don't always get specific information about why an email was blocked or delivered to the bulk folder.

Because filters are often a black box to email senders, they develop various theories about how or why spamfilters work the way they do. Some of the theories are true, but many of them are closer to myth than reality.

Senders that understand spam filters and the goals behind filtering can separate filtering myths from filtering truths, enabling them to create the most deliverable and effective emails for their target market.

In thinking about filters, it's important to understand the goals of the filter. There is no one-size-fits-all filter. Many filters are intended to focus on one specific type of unwanted or malicious email. Filters have different goals and therefore work on different parts of the email.

Filters aren't designed solely to block unsolicited bulk email, also known as spam. They're also useful for blocking malicious email, including virus infections and phishing emails. There are also user-specific filters, which can overrule any network filter. Mail that might otherwise go to the bulk folder will be delivered to the inbox.

Filters are different depending on the kind of receiver

When discussing filters, it is important to distinguish between filters on commercial ISPs and filters at businesses. Commercial ISPs cater to end users. Their business model dictates that they deliver mail that their users want to receive and block malicious mail or mail their users don't want. These filters take an average of the wantedness of an email, but allow individual end users to override the ISP decision.

Businesses, on the other hand, have email to further their own business goals. Many businesses allow employees to receive personal emails at their work address. However, they do not go out

of their way to facilitate personal mail delivery, so don't really care about things like opt-in status or permission. There are also workplace regulations on the types of email. Many businesses prohibit porn and filter against specific keywords to maintain a professional environment.

Businesses also may have different filters depending on their size. Large corporations often have mail systems that rival ISPs in population size and complexity. They may have employees maintaining filters and filtering software customized for the business needs and goals. Smaller and medium size businesses usually outsource their mail systems to commercial ISPs or spam filtering companies.

Steps in ISP spam filtering

Email delivery is a process, and filters interact with email during many different parts of that process. Email filters take a large mass of email and sort through it to separate the wheat from the chaff. While many of the specific tasks overlap, we can conceptually divide the process into 3 stages.

Stage 1: Should we accept this mail or not?

Stage 2: Should we deliver this mail to the inbox or the bulk folder?

Stage 3: How should we display this mail?

Each of these decisions is made at a different point in the process. The filter evaluates the status of the message according to a number of criteria specific to that stage. After the evaluation, the results are compiled and the server decides what to do with the mail. There are three decisions the receiver can make: this mail passes and can go to the next stage (or the inbox); this mail fails and should be rejected / discarded / bulk foldered; this mail is still unknown.

Stage 1: Should this mail be accepted?

The first stage of spam filtering happens when the sending mail server first contacts the recipient mail server. The receiving server must decide whether to accept the mail or not. At

this point, there is not much information known about the mail and there is no content. The one thing the receiver knows about the email is the IP address of the server sending the email.

Evaluation phase

Phase one of the acceptance stage is evaluating the sending IP. Filters check a number of criteria at this stage, all of which relate to the IP address sending the email.

Some ISPs do check domain reputation during this phase of email delivery. They look at domains (example.com, test.example.net), URLs (http://web.example.com/***) and email addresses (@mail.example.com) in the email. Most ISPs doing domain checks also resolve the IP address associated with a hostname and run checks against those IPs as well.

Blocklist check

Most receiving mailservers use some sort of IP-based blocklist. A blocklist is designed to prevent spam, virus or phishing emails from reaching the end user by preventing mail from the IP addresses on that list. There are a number of different types of blocklists that list varieties of harmful traffic. Some list IP addresses that look like they are infected with a botnet. Others list IP addresses that have sent spam as measured by different criteria like spamtrap hits or complaints. Still others list domains that are found in spam. Some just list domains or IP addresses from specific countries.

Checking an IP address against a blocklist is an extremely simple and cheap way to look at an email. But the action taken after the check varies — mailservers may block all mail from a listed IP address or simply tag the email as listed on a specific list.

Botnet check

Many viruses and botnet infections have characteristic behavior or configurations that distinguish them from legitimate mailservers. The receiving mailserver checks for these characteristics, and if the server looks like it is infected then mail is rejected.

Reputation check

When it comes to IP addresses, past performance is an indication of future results. If an IP address consistently delivers good mail, then it is very likely this new email is good, too. If an IP address consistently delivers bad mail, then it is very likely the new email is bad, too.

Of course, most IPs send a mix of good mail and bad mail, and the reputation falls somewhere in the grey area. This is where blocklist tagging comes in. If the mail is in the grey area, and is tagged on a blocklist, then it may be rejected. If the mail is in the grey area but not on a blocklist, it is passed onto content filters.

Some ISPs do check domain reputation during this phase of email delivery. They look at domains (example.com, test.example.net), URLs (http://web.example.com/***) and email addresses (@mail.example.com) in the email. Domains can be checked against internal or external blocklists.

Action phase

Depending on the answers to each test during the evaluation phase, the receiving mailserver decides to accept the mail, reject the mail or defer delivery for some period of time. In addition to the specifics of that particular email, the mailserver also evaluates its current state. If there is a heavy load on the server, more mail may be deferred than when there is a lower load on the server. In some cases load may get so high that deferrals are completely unrelated to any spam status of the email.

Every receiving mailserver has different thresholds, algorithms and specifics for what they will accept in an email. Many of them will not publish these specifics but will provide general recommendations for senders to resolve a block.

The mail server takes the action determined during the decision phase. At this point, the feedback that the sender receives can be one of three messages.

- We accept your email.
- We'd like you to wait and try later.
- We don't want this email.

Mail that passes the checks gets accepted into the mail server and is passed on to the next filtering stage.

Mail that fails all the checks is rejected.

Mail that is in a grey area can be tagged, accepted and passed onto filters, or deferred for later. When mail is deferred, the server can go collect more information in preparation for the next delivery attempt. Deferrals are not always consistent. Some deferrals are based on reputation alone; others are based on the load of the server and the reputation. Deferrals do not mean there is an actual problem with the mail. They may simply mean the server is overloaded. Delayed mail should be retried in a reasonable period of time.

Deferrals over long periods of time (hours or days) may indicate a problem with the IP reputation that should be addressed by the sender.

Stage 2: Where should we deliver the mail

After the server accepts the mail, content filters take over. All email entering this stage of filtering meets IP reputation thresholds and so IP reputation becomes a much less important factor in these decisions.

Content filters are expensive in terms of processing power. Many email servers are handling tens of thousands of emails a second. Since they reject many spam messages during the first stage, they don't have to commit so many resources during this stage.

Content-based filters look at a range of message components, from the actual text in the message, to the domains, to the IP addresses those domains and URLs point to. They look at the hidden structure of an email. They look at what's in the body of the message and what's in the headers. There isn't a single bit of a message that content filters ignore.

Some commercial filters even take a "fingerprint" of the email. They can compare the fingerprint with a database of known spam and known good mail. They can then determine how like spam the email is.

Content filters rarely use keywords to block mail. It's true: using "FREE!!!" in the subject line does not cause mail to be rejected out of hand. In the dim and distant past, there were keyword filters, but they were not very accurate. Eventually score-based filters replaced keyword-based filters.

Evaluation Phase

During the content evaluation phase, a filter runs a number of tests against the email. The results from the tests are assigned a numeric value. The types of tests range from the very simple to the very complex.

Some tests look for distinctive features from particular pieces of software. For instance, there was a piece of spamware that used a fake timezone value in message headers. Mail with that value was always spam. There was another type of spamware that forged a MS Outlook string. Mail containing this string was always spam.

Other tests look for features that are in both spam and non-spam. A lot of spam advertises diet pills, so mail mentioning weight loss may receive a positive score due to the word diet. Likewise, mail that mentions loans or business cards or Viagra may also receive a score simply for mentioning things often mentioned in spam.

Content filters also look at the domains and hostnames mentioned in the emails. These are evaluated based on the reputation of the domain and sometimes the IP address where the domain or hostname points. Domains and URLs have their own reputations separate from the reputation of the sending IP address.

Action Phase

Once all the tests have been run, the server has a numerical value for the email. This value is compared to the internal standard value and the email is either delivered to the inbox or the bulk folder. In very, very rare occasions the mailserver may determine this is a malicious enough email that it should be thrown away without notifying either the sender or the receiver. This is uncommon and should not happen to legitimate mailers.

During this phase, email delivery is determined for all recipients at an ISP or filter maintainer. Before the final delivery, however, individual user filters are consulted. Users are able to overrule these filtering decisions. They can add addresses to their address books, which results in the ISP delivering all mail from that address to the user's inbox. Users can also block mail from specific addresses, which results in the ISP not delivering that mail, or delivering it to the bulk folder.

Stage 3: How should we display the email

In this stage, the ISP determines whether or not it will display the email with all images or if they will add some annotation to the email. These filters act primarily based on the email content, the result of any authentication test, and enduser preferences.

Many ISPs are starting to display authentication results to the end user in the email client. Messages that pass authentication get a green check mark or other signal that the ISP has verified the email. Messages that fail authentication may receive a warning box or no comment at all.

Steps in business filtering

Businesses filter email in the same ways that ISPs do. But there are some differences that affect the delivery of an email.

It is important to distinguish between filters on commercial ISPs and filters at businesses. Commercial ISPs cater to end users. Their business model dictates that they deliver mail that their users want to receive and block malicious mail or mail their users don't want. These filters take an average of the wantedness of an email, but allow individual end users to override the ISP decision.

Businesses, on the other hand, have email to further their own business goals. Many businesses allow employees to receive personal emails at their work address. However, they do not go out of their way to facilitate personal mail delivery so don't consider things like opt-in status or permission. There are also workplace regulations on the types of email users may receive. Many businesses prohibit porn and filter against specific keywords to maintain a professional environment.

Businesses also utilize different filters depending on their size. Large corporations often have mail systems that rival ISPs in population size and complexity. They may have employees maintaining filters as well as filtering software customized for the business needs and goals. Smaller and medium size businesses usually outsource their mail systems to commercial ISPs or spam filtering companies.

Sometimes mail to business addresses is filtered because the business does not see business value in mail. Even opt-in mail can be filtered if business doesn't want employees distracted at work.

Conclusion

Senders who want to maximize delivery of key messages must understand how spam filters work and stay current with best practices for message development and delivery.