# Messaging, Malware and Mobile Anti-Abuse Working Group

# M3AAWG Email Anti-Abuse Product Evaluation Best Current Practices

**Updated**: March 2019 (August 2010)

Reference URL for this document: www.m3aawg.org/AntiAbuseProducts

Table of Contents

## Updated in this Version

The March 2019 version has been updated to note that email service providers and other organizations, in addition to internet service providers, may need to evaluate anti-abuse products for their networks. Several recommendations have been enhanced to consider the impact of evaluating products that will be affected by newer technology, such as cloud services, and by changes in anti-abuse products. New diagrams have also been added and portions of the text updated for clarity. An outdated survey has been deleted.

## Purpose

This document is intended to outline best common practices used when evaluating email anti-abuse products or services. Many enterprises and other messaging customers, not only ISPs and ESPs, are often tasked with evaluating some sort of anti-abuse solution but may incorporate different strategies and techniques when doing so. As a result of discussions within M$^3$AAWG, this document is a collection of best practice recommendations from various members of the anti-abuse community.

## Executive Summary

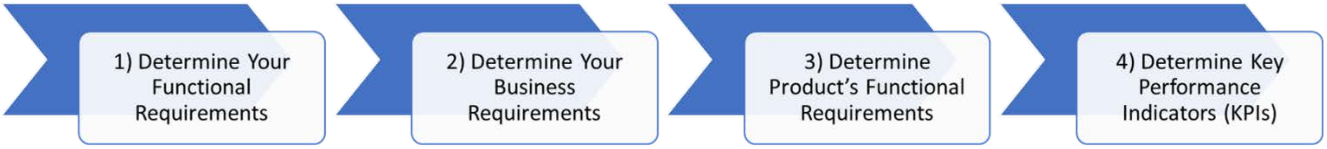
The methodologies malicious actors use in their attempts to deliver abusive mail shifts rapidly. As a result, testing anti-abuse technologies to find the most effective and accurate solutions is often problematic for ISPs and enterprises. Building on the experiences of M$^{3}$AAWG members, this document provides recommendations on processes and techniques to accurately determine a particular solution's effectiveness.

## Scope

This document's primary focus is on testing products or solutions that attempt to identify email messages as abusive: spam, malware, phishing, etc. As such, it applies mainly to products such as content filters that return a verdict or score based on an inspection of individual messages. Additionally, the techniques discussed can be applied to reputation-based products such as DNSBL lists (DNS-based Black Lists).

## Evaluation Planning

The following topics highlight recommended areas to consider when planning an anti-abuse evaluation.



1) Determine Your Functional Requirements

2) Determine Your Business Requirements

3) Determine Product's Functional Requirements

4) Determine Key Performance Indicators (KPIs)

### <u>1. Determine Your Functional Requirements</u>

Every tester will have their own set of functional requirements based on their existing environment, the type of product being tested, and the specific areas of concern that need to be addressed with a new solution. Some examples include:

- Hosted/SaaS versus appliance based
- Hardware architecture requirements (e.g. x86 vs. SPARC), if applicable
- Networking requirements
- Integration/compatibility with existing messaging platform
- False-positive (FP) and false-negative (FN) identification and mitigation methods (e.g., honeypot/trap monitoring, user feedback processing, etc.)
- Ability to take action on abusive messages (e.g., Can it quarantine to a "spam" folder? Can it tag the message using X-headers or subject header, etc.?)
- Reporting features and SIEM (Security Information and Event Management) capabilities, including troubleshooting and mail diagnostic toolsets
- Automated threat protection and intelligence

## 2. Determine Your Business Requirements

Business requirements may not be necessary in order to conduct a technical evaluation but it can be beneficial to understand whether the vendor is agreeable to such requirements prior to testing. These requirements may include:

- Support services
- SLAs (Service Level Agreements) for:
  - Response times regarding service-affected issues
  - Filtering accuracy issues
  - Filtering accuracy thresholds
- Costs, not only for the product itself, but including:
  - Potential integration costs into an existing platform
  - Operational support costs
  - Training costs
- Business support/help desk impact – How will the product affect support and administrative staff workload?
  - Initial onboarding costs
  - Ongoing maintenance and monitoring costs
  - Differences in workload if replacing an existing anti-abuse solution
  - Training needs
- Vendor reputation should also be considered. Financial stability, size and industry reputation are areas to research when vetting a product and its vendor.
- Privacy concerns – Does the vendor and their product conform to your organization's privacy policies, both during the evaluation and in a potential full production deployment?
  - With the proliferation of cloud services, it is important to identify your organizational needs around data privacy and the implications on what are, and are not, acceptable environments and inherited policies under which your data may reside.
- Compliance – Does the vendor and their product meet your compliance requirements? Is the vendor committed to European Union's GDPR (General Data Protection Regulation) compliance?
- Service and vendor availability
  - Some enterprises require redundancy in services and availability, especially for core systems.
  - Compatibility with existing solutions is important to understand.
  - It is also important to evaluate whether you want a single provider for multiple security features (to reduce complexity/supportability problems) or multiple providers to add in redundancy.

## 3. Determine the Product's Functional Requirements

Understanding the functional requirements of the product will help in preparing the testing environment for its introduction. For example, some products may require:

- Outbound firewall ports open for update retrieval or other purposes
- Vendor's support personnel to access the system remotely

These requirements need to be known in preparation for testing and also to determine whether or not they conflict with any corporate or security policies.

### 4. Determine Key Performance Indicators (KPIs)

When planning an evaluation, the Key Performance Indicators (KPIs) of the overall effectiveness of a product need to be determined. The following is a list of KPIs related to anti-abuse products:

- **Filtering Percentage (aka "catch-rate")** - The percentage of total traffic which the product identifies as abusive.
- **Filtering Accuracy** – A percentage that represents how much traffic was correctly identified as either legitimate or abusive. This is derived from the false-negative and false-positive rates defined below and may vary by type of attack or environment where the solution is deployed.
- **False-Negative Rate** - The amount of abusive traffic that the product does not catch.
- **False-Positive Rate** - The amount of legitimate traffic which is incorrectly marked as abusive.
- **Reaction Time** – How fast does the product react and adapt to new attacks?
- **Stability/Fault Tolerance** – How do potential product failures or issues impact overall service?
- **Message Throughput** - The amount of traffic the product can process at a given time (i.e., messages per second).

This list is not meant to be definitive but merely include some recommended areas of focus during an evaluation.

## Conducting the Evaluation

An evaluation will generally consist of overall functional testing (which may or may not include load testing, depending on the type of product), followed by testing of filtering accuracy. Accuracy testing is by far the more difficult of the two to perform but usually is of most interest when conducting a test.

### Functional Testing

Once the KPIs and requirements have been determined, an initial evaluation of the product in a test environment should follow. This initial phase should mainly focus on confirming the functional requirements that were set forth prior to the evaluation and preparing the product for the upcoming accuracy testing. During this time, working with the vendor to configure the product optimally is vital. This will help to ensure that the accuracy tests conducted later produce reliable results.

Since this portion of the evaluation is not particularly focused on filtering accuracy, most methods of passing test traffic through the product should be acceptable to gauge overall functionality. It is recommended that a message corpus be used that has varied content within the body, including headers and varied message sizes. A mixture of HTML content, images and attachments should be included. A message corpus taken from a sample of production traffic is ideal, where possible.

The following questions should be answered during functional testing:

- Does the product satisfy the core functional requirements (aside from accuracy)?
- Do the tester and vendor agree that the product's configuration is optimal?
- Does the product satisfy required load and throughput requirements?

### Load Testing

Load testing to determine message throughput and system utilization is necessary when testing products such as appliances or software-based content scanners. A proper load testing tool that can generate sufficient volume should be used. Most load generation utilities that support SMTP should be sufficient. When performing load tests, use similar message samples as those in the functional testing; that is, messages with varying content and size.

## Accuracy Testing

This is the most challenging part of the evaluation, mainly due to the limitations inherent in most testing environments. Since labs normally do not see the same traffic that flows through the production environment, it is difficult to ascertain the accuracy of a product in the lab. Therefore, some choose to gauge accuracy using partial or phased production trials. Others opt for techniques that divert production message traffic into a lab environment. Each has its advantages and disadvantages which will be discussed further.

When deciding which method to use when testing for accuracy, there are two main areas to consider:

### Corporate Policies

When outlining your testing process, it is vital to understand and incorporate your organization's policies on production system usage for evaluations. Your organization's privacy policies regarding message traffic used for testing are also important.

### The Type of Product Being Tested

The type of product being tested and the general technology it uses can determine the most appropriate environment in which to test it. For example, a heuristic-based filter which mainly looks at message content can produce accurate results in a lab environment for certain KPIs. However, products that act within the SMTP session or otherwise depend on how the sending MTA behaves (e.g., edge appliances or grey-listing products) may only yield accurate results in a production environment.

## Production Trials

The following lists the advantages and disadvantages of conducting production trials to gauge accuracy.

### Advantages

- Can react directly to the sending MTA's behavior
- Can gauge effectiveness in real-time
- True end-user feedback for false-negatives/false-positives
- Gaming by spammers can be observed

### Disadvantages

- Evaluating products in the production environment risks undesirable results that might affect customer experience, such as high FPs/FNs, unexpected service outages, etc.

### Precautions When Conducting Production Trials

When conducting trials in production, it is imperative to do so in a manner that minimizes adverse effects on end-users. Therefore, the introduction of the test product should be designed in such a way that the entire incoming message stream is not exposed. It should also be designed so that the product can only take action on messages that are destined for accounts that are participating in the trials (honeypots/spam traps, test accounts, etc.).

### Trap Analysis

Assuming that the traffic to spam traps is all spam, false-negative accuracy can be gauged by enabling the product to filter traffic destined for these accounts and then examining the results. It is recommended that the product be configured to mark the message with an X-header, or perhaps in the subject header, showing the result of the test product's filtering, if possible. This way the message set can be analyzed to determine the false-negative rate. If marking the message is not possible, then the test should be configured to drop the message if it is deemed spam by the product. The results can be determined by examining the remaining traffic. This should also produce similar false-negative results.

One drawback to spam trap analysis may be the volume that spam traps normally receive. If the volume of traffic these accounts normally receive is too small to be statistically relevant, they may not produce an accurate view of false-negative accuracy. In this case, it will be necessary to use data from test user account feedback (explained in the next point below below) to gauge this. Another drawback is that because the accounts are not true accounts, sophisticated attacks will likely not be targeted at the honeypots – the evaluation may give an unrealistic view of the solution's accuracy.

User Acceptance Testing Analysis

The next step is to begin filtering traffic destined to a set of end-users knowingly participating in the evaluation. The test product would be configured to act on messages that are destined to these users only. Similar X-header insertion as described earlier would be used here, with the users providing feedback on both false-positives and false-negatives. It is recommended that the users not be made aware of a message's disposition (i.e., whether it was marked as abusive or not), which could potentially impact their feedback. Their feedback is then used to determine the false-positive and false-negative rates.

Partial Production Deployment

The final production trial would be a partial deployment to the general end-user base. How this is accomplished depends on the particular production platform. For example, if there are two or more MX records that represent two or more ingress points, the test product can be used to filter one while the others are used as a control group. If this is not possible, then deploying the product in production for a set duration of time can be a plausible alternative. However, this exposes all end-users to the new product, which may be undesirable, but it would give the best indication of how well the product performs since it would provide the most false-positive and false-negative feedback. Whichever means is used, the important consideration is that the product is exposed to a large enough set of end-users to provide meaningful numbers in terms of false-positives and false-negatives.

## Lab Testing for Accuracy

The following lists the advantages and disadvantages of accuracy testing in a lab environment.

Advantages

- Can gauge effectiveness in real-time
- Does not risk end-user experience
- Allows for simultaneous testing of multiple products

Disadvantages

- False-positive/false-negative accuracy more difficult to gauge
- Cannot react to the sending MTA's behavior
- Cannot detect gaming by spammers

Lab testing does have some limitations in determining accuracy compared to production testing. It also cannot produce accurate results from products that depend on the behavior of the sending MTA, but it will likely be the method of choice for those environments that are unable to conduct production testing for policy reasons. It also allows for the simultaneous testing of multiple products using the same message stream, which is vital for accurate comparisons between competing products. See Diagram 1 below that shows how this can be accomplished.

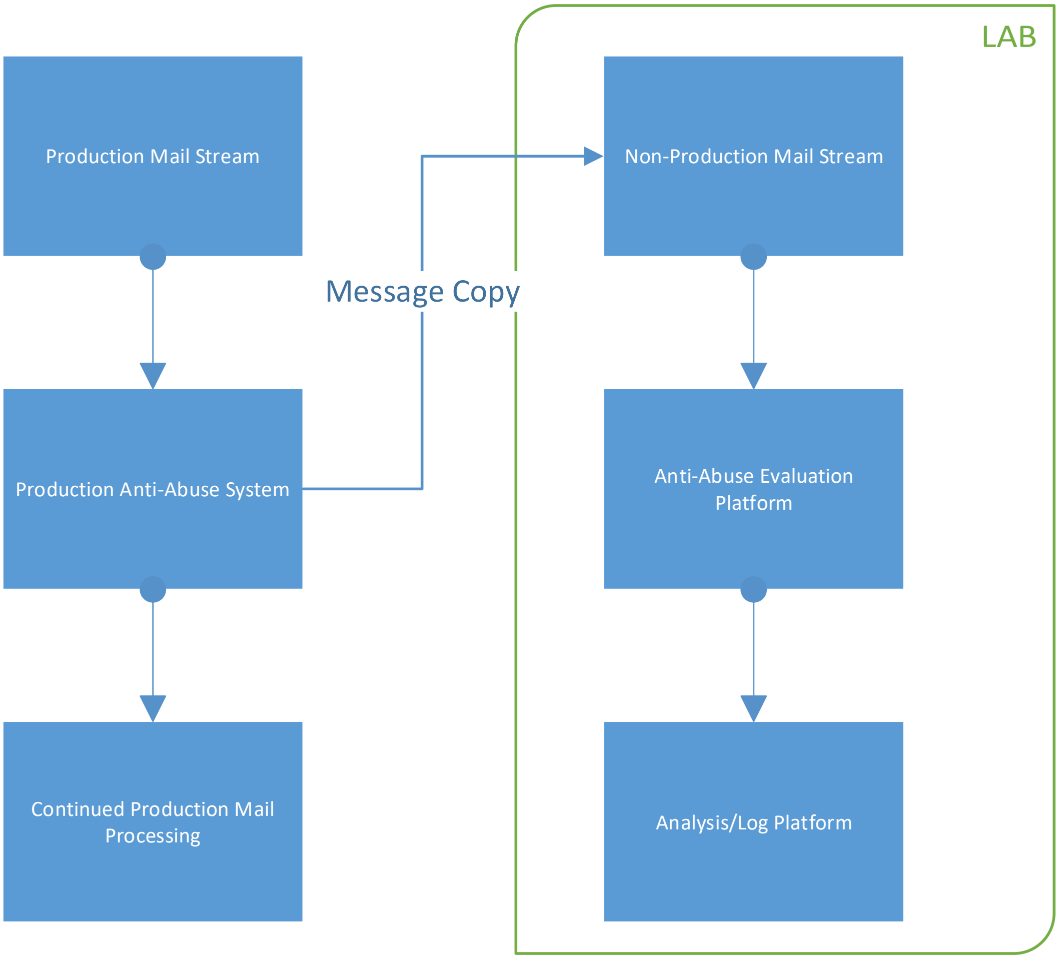
Live Traffic Streams

Lab testing should use live traffic, or as close to real-time traffic as possible, since most anti-abuse products today provide the most reliable results when processing more recent traffic. This is accomplished by forking a copy of production traffic into the lab.

Sending Production Message Copies into the Lab

This technique involves configuring a production server to send a copy of each message it receives into the lab environment for testing. The following diagram shows how this can be done.

## Diagram 1 - Production to Lab Flow


LAB
Production Mail Stream
Non-Production Mail Stream
Message Copy
Production Anti-Abuse System
Anti-Abuse Evaluation Platform
Continued Production Mail Processing
Analysis/Log Platform

As shown above, the message copies are sent to the lab platform, which then processes the message copy through its own, separate system. A simple way to test would be to mark the message by inserting a X-header value with the system's verdict for later review. It is recommended that any current anti-abuse filters being used in production be a part of this testing. This allows for easy comparisons with the test products.

Preserving the Message Source IP

**XCLIENT** (see http://www.postfix.com/XCLIENT_README.html)
One potential issue with sending message copies to the lab is that the IP of the production server is sourcing all the traffic to the lab. There is a SMTP extension called XCLIENT that can be used to pass the source IP of the original sender into the lab. The lab server then treats this IP as the source IP instead of the production server. This allows testing of products that require the message's source IP in order to make decisions. Both the production server that is sending messages to the lab and the lab server receiving the messages must support the XCLIENT command.

Adding Source IP to Message Headers

If XCLIENT is not an option in your particular MTA, an alternative method is to have the production MTA add a specific X-Header with the source IP of the message. Then the receiving MTA in the lab can be configured to use the IP listed in this header as the source IP of the message during processing.

Precautions

It is important that care be taken when configuring the production system that will perform the message copying to the lab environment. It should be configured so that it does not cause undesired effects if it encounters delivery issues to the lab. Issues such as generating bounce messages or NDRs and allowing potential queuing of lab copies that would affect production service should be avoided.

Message Corpus Testing

When possible, use live streams of known traffic types and pass this traffic into the lab environment for analysis. For example, here are some sources of spam that can be fed into the lab:

- Honeypot traffic
- Traffic from IPs on highly-trusted DNSBLs
- Messages reported as spam by end-users

These sources usually are all considered spam and can help in determining false-negative rates.

Any known good traffic streams can also be sent to the lab to help determine false-positive rates. The following lists some sources of mostly good traffic:

- Messages or IPs whitelisted in production
- Messages reported as not spam by end-users
- Messages from accredited sources

Manual Review

Depending on the confidence level of the quality of these streams, a manual review of messages may still be necessary. This involves reviewing the contents of the message and deciding whether it is abusive or not. This can be a time-consuming process depending on the volume of mail that was scanned. Therefore, if more than one product is being evaluated concurrently, the messages for which all products agreed can be omitted from manual review. In other words:

- If all products marked a message as spam, it can be considered spam.
- If all products marked a message as legitimate, it can be considered legitimate.
- If there are disagreements between products for a message, the message should be manually inspected.

Inspecting only those messages where products disagreed can still result in a large number of messages to review. To help with this, some of those who previously conducted evaluations have developed simple in-house tools that displayed a message within a web browser with two choices allowing the tester to mark it as abusive or not. For example, if a tester marked a message as spam with this tool but the product found the message to be legit, the tool would record it as a false-negative for that product. These results are tracked for all products being tested. After all messages have been reviewed, the resulting false-negative and false-positive data is analyzed.

Calculating Accuracy

Once the false-negative and false-positive numbers are determined from a set of messages, the following equation can be applied to calculate accuracy:

*Accuracy % = 100*(((# Messages Scanned) – (# FP + # FN)) / (# Messages Scanned))*

For example, if 100 messages were scanned, and a product produced three false-positives and seven false-negatives, the resulting filtering accuracy would be 90%. Note that this equation assumes that FNs and FPs equally impact a product's overall accuracy. However, their levels of importance may differ based on the tester's particular criteria. In this case, it is recommended that FP and FN percentages be looked at individually to gauge overall accuracy rather than calculating a single percentage value. Depending on the disposition of the mail, some organizations may be more or less tolerant of false-positives – for example, some are more willing to trade off false-positives for a malware catch than for a spam catch.

## Concluding the Evaluation

After the evaluation is complete you should analyze how each product performed against the various metrics and functional requirements defined at the beginning of the process, with priorities assigned to all of those KPI metrics and requirements. Evaluate against any hard thresholds that the business has for metrics that cannot be exceeded (e.g., utilization, costs, throughput). For products that meet your requirements, the company should weigh the effectiveness numbers. Then the following Decision Matrix can be used to summarize how all the products performed overall.

### Decision Matrix

It is recommended that a Decision Matrix be used to compare the results of the various products. A sample matrix looks like this:

| Decision Matrix | | Solution A | | Solution B | | Solution C | | Solution D | |
|---|---|---|---|---|---|---|---|---|---|
| **Criteria** | **Weight** | **Rating** | **Score** | **Rating** | **Score** | **Rating** | **Score** | **Rating** | **Score** |
| Filtering Percentage | 15 | 4 | 60 | 4 | 60 | 4 | 60 | 4 | 60 |
| Filtering Accuracy | 15 | 3 | 45 | 3 | 45 | 2 | 30 | 2 | 30 |
| False-Positive Rate | 12 | 4 | 48 | 4 | 48 | 3 | 36 | 3 | 36 |
| False-Negative Rate | 12 | 4 | 48 | 1 | 12 | 3 | 36 | 3 | 36 |
| System Resource Utilization | 11 | 4 | 44 | 3 | 33 | 3 | 33 | 4 | 44 |
| Throughput | 10 | 4 | 40 | 2 | 20 | 3 | 30 | 4 | 40 |
| Costs | 8 | 2 | 16 | 4 | 32 | 4 | 32 | 3 | 24 |
| Help Desk/Support Impact | 5 | 4 | 20 | 3 | 15 | 3 | 15 | 4 | 20 |
| Vendor Reputation | 4 | 4 | 16 | 4 | 16 | 4 | 16 | 4 | 16 |
| **Total** | 100 | **36** | **361** | **32** | **313** | **33** | **320** | **35** | **338** |

The criteria listed here are only examples, taken from the KPIs and business considerations mentioned earlier. The actual criteria used will depend on what KPIs are deemed most important by the evaluator. Each of the criteria is assigned a "Weight" which signifies its importance in the decision-making process. (The weights listed here are again examples only. The evaluator should assign proper weights as appropriate to their needs.) The "Rating" column is a value assigned based on a scale of how well the product performed for that particular criteria. The above example uses a 1-4 scale, 1 being bad and 4 being good.

The "Score" columns are the result of multiplying the "Rating" value by the "Weight." The individual scores are then added for all of the criteria to produce a final score for the product. This final score can be used to determine which solution (or solutions) performed the best overall. In this example, the best overall product was Solution A, followed by D, then C, and then B.

## Conclusion

Anti-abuse product evaluations need to be tailored to the needs of the organization and its specific business and operational requirements. It takes planning and careful preparation to complete an extensive analysis of one or more products. This paper looks at the options and provides suggestions based on expertise from M$^3$AAWG members across multiple aspects of the anti-abuse community.

As with all documents that we publish, please check the M$^3$AAWG website (www.m3aawg.org) for updates.

M3AAWG030-2019