Best Practices for Scanning Creative for Malware

This document describes a proposed set of responsibilities and options a seller has when receiving ad tags and ad creative, with regards to scanning this content for malware.

Version 1.0

Released October 2016

About the Anti-Malware Working Group

The rapid and continued growth of the digital advertising ecosystem has allowed unscrupulous actors to take advantage of parties at all points along the chain to insert malicious advertising, which only serves to dilute trust and harm the health of the landscape.

The mission of the Anti-Malware Working Group is to contribute to the healthiest possible ecosystem by improving trust, transparency, and accountability among all parties involved by developing the tools, standards, and technologies that enable the elimination of malware.

About the Trustworthy Accountability Group

The Trustworthy Accountability Group (TAG) is a first-of-its-kind, cross-industry accountability program fighting criminal activity across the digital advertising supply chain. TAG works collaboratively with companies throughout the supply chain in four areas critical to the continued growth and development of the $50 billion digital advertising industry:

- Eliminating Fraud
- Combatting Malware
- Fighting Internet Piracy
- Promoting Transparency

A joint marketing-media industry program, TAG was created by the American Association of Advertising Agencies (4A's), Association of National Advertisers (ANA), and Interactive Advertising Bureau (IAB).
To learn more about the Trustworthy Accountability Group, please visit www.tagtoday.net.

Contacts

Brendan Riordan-Butterworth - Senior Director, Technical Standards
brendan@iab.com

Jamie O'Donnell, Manager, Compliance Programs
jamie@tagtoday.net

Executive Summary

This is a foundational piece for the industry establishing better coordinated and robust malware detection and prevention practices.

Contents

Scope

These guidelines are intended to cover detection of malware originating in, and spread by, online advertising. The guidelines are applicable to all environments in which digital advertising are displayed.

These guidelines are not intended to cover other issues, such as click fraud or piracy. Nor are these guidelines intended to cover inappropriate or illegal content in ads.

This document is principally applicable to those organizations involved in the digital advertising ecosystem, including but not limited to: DSPs, Agencies, Trading Desks, Technology Providers (Ad Servers, Hosting Platforms, Scanning Services, …), SSPs, and Exchanges.

Principles

Scanning is Required

All ads and landing pages require scanning against malware by using either in-house and/or a reputable 3rd party service

- Scanning should incorporate updated blacklists that account for new threats
- Scanning should strive to detect and recognize threats that are at times hidden but still exposes users to malware (cloaking)
- Scanning of ads and landing pages may be performed asynchronously, but it is recommended to precede a landing page scan with an appropriate ad request.

Scanning Precedes Delivery to Consumers

Ads and landing pages should be scanned, near real-time and preferably prior to first user exposure

- If you host the creative, you should do this during the campaign setup process.
- If you do not host the creative, you should make a best effort to scan ahead of the first user exposure.

Regular Rescanning

Ads and landing pages should be re-scanned with appropriate resources in proportion to the user exposure, technologies, and partner confidence

- Scanning an ad only once is likely insufficient and should be re-scanned periodically.
- There are no absolute numbers required for scans.
    i. It should be everyone's goal to reach a high confidence that malware is not present by scanning a mathematically appropriate amount, adjusted based on factors outlined in this document. For instance, ads with millions of impressions per day may need hourly scanning, while ads with a hundred impressions per day may need weekly scanning.
    ii. All participants must use commercially reasonable and best efforts to provide acceptable coverage.

Malware Delivery Methods

Table identifies what modes of delivery should be caught.

Type	Recommend Detection	Details
Creative Has Malware	Yes	The creative payload delivers a malicious program directly.
Landing Page has Malware that installs without user interaction	Yes	After the user interacts with an ad and is redirected to the expected landing page, malware is installed automatically (eg, customized drive-by download).
Landing Page has Malware that requires user interaction to install malware	No	After the user interacts with an ad and is redirected to the expected landing page, they are prompted to install malware (eg, fake “outdated plugin” prompt).
Landing Page promotes a link that is Malware	No	After the user interacts with an ad and is redirected to the expected landing page, they are provided with a link to another page that contains malware.

Responsibilities

Participants in the supply chain pass along and modify the Ad Tags while not necessarily handling the Ad Creative (see Glossary). Depending on whether an entity has access to the full ad or only parts of it, their ability to scan varies. As such, different scanning responsibilities are recommended.

Ad Creative Handlers

For entities with access to the creative

- Creatives and landing pages must be screened for malware, as specified in The Principles

Ad Tag Handlers

For entities with access to the tag

- Tags, all intermediate calls, and landing pages must be screened for malware, as specified in The Principles

All Parties

For all entities, with access to the creative or tags

- Only after screening can the entity assert the creative is potentially malware-free; Entity may continue to assert this while re-screening, as specified in The Principles
- If a new threat is identified, the entity may only continue to assert a creative is potentially malware-free if re-screened inspecting for that threat.

Scanning Factors

A large variety of factors may influence the best cadence of scanning to achieve effective malware detection. They are enumerated following in order to inform the development of specific scanning practices by each individual company.

Demand Type

The breadth of clients that an ad may be delivered to will affect scan frequency. When ads may be delivered across multiple types of connections to a variety of devices, scans should be increased to reflect this diversity. Factors to consider include:

- Whether malware may deliver to specific ranges of IPs that represent public Wifi or cellular connections.
- Whether malware may deliver to specific cookie data.
- Whether malware may deliver only over HTTPS or non-HTTPS connections.
- Whether malware may deliver only to certain geo-locations.
- Whether malware may deliver only to a subset of User Agents: Browser/OS and App/Mobile Web (mobile/tablet, iOS/Android) combinations.

Frequency

The number of ad impression and ad click data should affect scan frequency. Multiple scans should be performed when an ad is being delivered more often, and other factors to consider include:

- Absolute number of impressions or other common KPIs. This describes a general exposure risk.
- Acceleration of impressions or spend. While mid-campaign changes are often innocuous, they may reflect a malware activation.
- Changes in impression targeting. Fine tuning a campaign's targeting can be a normal part of the effort, but also might represent the activation of malware delivery.
- Changes in Technology. A shift from static creative to dynamic creative with external resources, introduction of a redirect or number of redirects in a chain may all indicate an increased need of scanning.

Rescan Factors

Scanning ad creative and ad tags directly is an ongoing cost. In order to achieve an effective and commercially reasonable scanning cadence, additional data points can be used to inform the rescanning interval:

- Initial scan results. The complexity of the ad encountered during the initial scan, as evaluated by the number of URLs, number and size of reference JavaScript files, and so on, can influence the frequency of follow-up scans.
- Sub-element errors. If during any scan there are technical issues in making requests (like 404 errors, DNS errors, server timeouts, and so on), rescanning cadence should be increased.
- Domain location. The physical location of the domains referenced in the ad tag should be considered. Hosting at a well known co-location facility presents a different risk profile than resources hosted by fast fluxing DNS on residential or proxy IPs.
- Domain and IP ownership information. The ownership information about domains and their associated IPs referenced in the ad tag should be considered. Domain WHOIS data from ICANN accredited registrars, and IP Whois data from ARIN, RIPE, APNIC, and so on should be considered here.
- Partner confidence. The current length of partnerships with buyers should be considered in rescanning cadence, as well as the assertions made by and track record of the buyer with regards to scanning and detecting malware themselves.
- Technology. Some technologies are intrinsically higher risk since they they contain an increased potential for delivering malware (like SWF files), or because they are different for each delivery (like rotating tags).

Cost / Value

Not all scans are equal cost. Frequency of higher cost scans can be reserved for the coverage commensurate with the value gained (and insuring that other serving metrics are not impacted). The order of costs are:

1. Simple Rescoring. Analysis of previously collected scan information against updated database of threats.
2. Static Analysis of Creative. Analysis of previously collected creative against updated database of threats.
3. Low- to Full-feature scans. Actual scanning, from simple requests for the ad tags through to scans made from emulated browsers that interact with ads and request landing pages.
4. Scans from consumer IP spaces. Actual scanning, but instead of from a datacenter, the scans originate from known consumer IP addresses.
5. Scans from various mobile networks. Actual scanning, but instead of from a datacenter, the scans originate from known mobile network IP addresses.

Hosting Ad Tags and Ad Creative

In our opinion, the strongest current signals for risk of demand is whether the creative is self-hosted by the demand partner, or if it's externally hosted. The next strongest risk signal is whether or not the creative is active (JavaScript, java, flash, etc.) or static (png, jpeg, gif, txt, etc.).

Creative Risk Levels

Sorted high to low risk. Security thinking treats any 3rd party resource as a potential threat, so understanding what a malicious buyer could deliver if they chose to be an attacker.

Active creative hosted remotely

- Buyer can subvert the supply chain at multiple levels, and is difficult to detect or interdict.

Static creative hosted remotely

- Buyer can subvert the supply chain at multiple levels, and is easier to detect than active content (type/mime mismatch), but still difficult to interdict.

Active creative hosted locally

- Buyer can still potentially make remote calls out to other remote resources, if buyer's own processes are insufficient.
- Improper scrubbing on the part of the buyer can easily put this in the same risk category as active content hosted remotely.

Static creative hosted locally

- Lowest risk
- Buyer has really only one chance to get their payload into the system
- Scanning static creatives for "bad things" is more well defined than active creatives.

Recommendations for handling creatives

These recommendations should be considered the minimum for acceptance of their demand into the supply chain, and are ordered by risk.

Self serve buyers should structure their buys such that the majority of their inventory fits into the lowest risk categories, when possible.

Active content hosted remotely

Highest risk. JavaScript on a system that you have no control over.

- Scan ad tags with robust malware scanning.
    - A well supported scanning system with a reputation for success is recommended.
    - Reputation will impact the long-term viability of any homegrown scanning technology.
    - Rescan more frequently, since this is active content.
- On initial creative ingest, fetch all remote resources, and do the following:
    - Block a creative that involves eval()
    - Block a creative that has remote resources within remote resources.
    - Block a creative that appears to have excessive base-encoded data and also contains a call to a decoder function.
    - Determine if a checksum or other digital signature can help speed subsequent rescans, since these will be frequent.
- Ingest verifiable proof of scanning by the hosting party.
    - If the hosting party is able to provide verifiable and trusted proof of their own malware scanning, you may be able to reduce your scanning interval.

Static content hosted remotely

High risk. Creative on a system you have no control over.

- Scan creative with robust malware scanning.
    - Homegrown or 3rd Party malware scanning tools are acceptable.
    - Reputation will impact the long-term viability of any homegrown scanning technology.
    - Rescan at a reasonable interval, since the static content can be changed (HTTP 302 and so on).
        - Deep scan on creative change. Use checksum or full bit comparison.

- Ingest verifiable proof of scanning by the hosting party.
    - If the hosting party is able to provide verifiable and trusted proof of their own malware scanning, you may be able to reduce your scanning interval.
- On initial creative ingest, attempt to check the following:
    - Minimum and maximum file size.
    - Able to correctly resize the image without throwing an exception.

Active content hosted locally

Moderate risk. JavaScript on a system that you have control over.

- On initial creative ingest:
    - Ensure creative is well formed. How this is performed technically is dependant on the format of the creative:
        - Flash presents a small technical hurdle.
        - Java applets should be immediately suspect, given their high power over the local system, and relative rarity of their appearance.
        - Block calls to eval()
        - Potentially block any heavily encoded payloads, especially double and triple encoded.
    - Determine if there are any remote systems called inside this creative
        - If there are, assess reputation of those endpoints.
- Subject the creative to long-term remote scanning (connecting as a client would), whether by a 3rd party or by a homegrown solution.
- Consider creating verifiable proof of scanning.
    - As the hosting party, you may be able to provide verifiable and trusted proof of your own malware scanning, to help reduce the scanning interval or partners that you are buying from.

Static content hosted locally

Low risk. Creative on a system that you have control over. “Directly Hosted Static Content”

- On initial creative ingest, check the following:
    - Ensure creative is well formed. How this is performed technically is dependant on the format of the creative.
- Monitor the local file for changes on a routine basis, and re-scan on change.
    - New upload
    - Potentially compromised system changing the creative outside the application.

Conclusion

This document has presented a set of recommendations for scanning malware - what to scan, how often to scan, and what factors to take into account when determining frequency and depth of scans.

Glossary

Ad - The superset of both creatives and tags. Different participants in the supply chain may have access to only creatives or only tags, so this is shorthand for both.
Creative - The actual payload that, when delivered to a web browser or other client, displays a message to the consumer.
Tag - The javascript or HTML code that indicates to the web browser or other client the location of the creative. Tags may also indicate the location of other tags.
Landing Page - After the consumer interacts (ie. a “click through”), this is the final site or app destination.
Scan - Analysis by means of static examination, virtual execution/emulation, or manual rendering of a creative or tag. This can and should also include any actions caused from user interaction, such as a click, and should also include malware detection on the Landing Page.
Scan Frequency - The regularity of which an ad is evaluated, scanned, or rendered for the purpose of compliance checks.
Malware - Any malicious software installed on a computer or device (ie. phone, tablet), without user consent. This can include (but not limited to) spyware, worms, bots, viruses or adware.

Additionally, see the TAG Fraud Taxonomy for more ad related terms.