A Machine Learning Method to Block Ads Based on Local Browser Behavior

Originally published at https://www.unite.ai/a-machine-learning-method-to-block-ads-based-on-local-browser-behavior/

Researchers in Switzerland and the US have devised a new machine learning approach to the detection of website advertising material that's based on the way such material interacts with the browser, rather than by analyzing its content or network behavior – two approaches which have proved ineffective in the long term in the face of CNAME cloaking (see below).

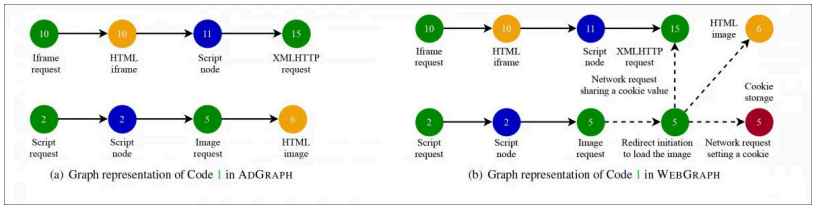
Dubbed WebGraph, the framework uses a graph-based AI ad-blocking approach to detect promotional content by concentrating on such essential activities of network advertising – including telemetry attempts and local browser storage – that the only effective evasion technique would be to not conduct these activities.

Though previous approaches have achieved slightly higher detection rates than WebGraph, all of them are prone to evasive techniques, while WebGraph is able to approach 100% integrity in the face of adversarial responses, including more sophisticated hypothesized responses that may emerge in the face of this novel ad-blocking method.

The paper is led by two researchers from the Swiss Federal Institute of Technology, in concert with researchers from University of California, Davis and the University of Iowa.

Beyond AdGraph

The work is a development from a 2020 research initiative with Brave browser called AdGraph, which featured two of the researchers from the new paper.



Comparison of AdGraph vs. WebGraph, with dotted lines representing architectural innovations on the previous approach. Source: https://arxiv.org/pdf/2107.1130

AdGraph relies on (ad) content features, derived from analysis of URLs, as a key to the detection of commercial material. However, these features represent a single potential point of failure for adversaries seeking to detect the presence of ad-detection systems, and formulating methods to obviate them. This reliance on content *properties* makes AdGraph essentially a mechanized version of manually-curated filter lists-based approaches, sharing their weaknesses.

CNAME Cloaking

Material originating from a website's own domain falls into a 'trusted' category, insofar as the domain itself is trusted. For a high authority website, there's a valuable premium in running advertising campaigns that feature material that *appears* to be hosted by the authority site itself, since such advertising is immune to filter-based ad-blocking lists, and even to the 2020 AdGraph approach.

However, custom campaigns are difficult to negotiate, expensive to implement, and run contrary to the core principles of the network advertising model developed over the last 25 years, where a third-party platform inserts code directly into the host site, usually 'auctioning' the advertising slot in microseconds based on keyword desirability and various other factors.

Since nearly all ad-blocking systems key on third-party material in web pages (i.e. elements which are hosted on 'alien' domains), advertisers have been fighting back with CNAME cloaking techniques over the last five years. CNAME cloaking deceives trackers into believing that a subdomain of the host site (i.e. information.example.com instead of example.com) is a genuine adjunct to the site, when it fact it is a proxy ad-serving mechanism arranged with third party advertising providers.

In March of 2021 one study revealed that CNAME cloaking incidents increased by 22% between 2018 and 2020, with nearly 10% of Tranco's top 10,000 websites employing a minimum of one CNAME-based tracker by October of 2020.

Discounting Trust in URLs

CNAME deception techniques involve manipulation of URLs involved in the ad-serving process. Any ad-blocking system that trusts the URL chain will be subject to manipulation and evasion. Therefore WebGraph randomly changes the supplied URLs in a process (including query strings, number of parameters and parameter names), looking for patterns of use rather than specific banned or accepted URLs.

The system has to consider two common configurations in an ad-serving architecture: one, where the host is colluding directly with the advertiser; and a second (more common) scenario where the advertiser provides limited cooperation due to the need to protect itself against manipulation by its clients.

In list-based approaches, including AdGraph, successful URL manipulation by the ad-serving system is almost a complete victory, ascribing 'local' provenance to the ad, and therefore evading nearly all attempts to systematically block advertising content.

What's left, by way of signature? WebGraph focuses instead on advertising systems' need to share information by various semi-obfuscated means, such as web trackers, communications between iframes and web 'listeners', which are constantly polling the live state of the host page for activity that is meaningful in terms of web-metrics for the ad. Such activity includes the storage of variables in cookies or HTML5-based local storage.

WebGraph uses Mozilla's Web Privacy Measurement (OpenWPM framework) to track such activity in Firefox. It captures all activity at the JavaScript layer, and all outgoing network requests, and their responses, at the network layer.

This additional scrutiny introduces new 'information flow' edges to the graph network previously proposed by AdGraph, allowing WebGraph to explicitly record and quantify information sharing patterns based on local activity, and regardless of origin and destination URLs for telemetry or other kinds of internecine communications in ad-serving systems.

Results

The researchers used an extended version of OpenWPM to systematically crawl 10,000 websites taken from the Alexa top 100,000 sites, and a random sample of 9,000 sites ranked between 1k-100k, storing their graph representations before passing the results to a decision tree classifier modeled on AdGraph's original design, and using popular ad-filter lists as ground truth. In this way, a dataset was constructed for training of the core model.

The system achieved comparable results to AdGraph, with 92.33% accuracy. However, the new system's resilience to adversarial resistance rises from an almost complete failure rate for AdGraph to just 8% susceptibility under WebGraph.

Future Directions

The paper contends that ad networks would need to notably re-architecture their systems in order to evade detection in the face of the WebGraph approach, and suggest that such changes would necessitate a review of the currently circumspect trust relationship between third-party advertisers and the host sites on which their ads appear.

The paper also notes that WebGraph does not take account of stateless tracking techniques such as browser fingerprinting (via the Canvas element), which make use of APIs that the system does not currently monitor. The researchers suggest that WebGraph can be extended in the future to account also for those kinds of interactions and local storage signifiers.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai