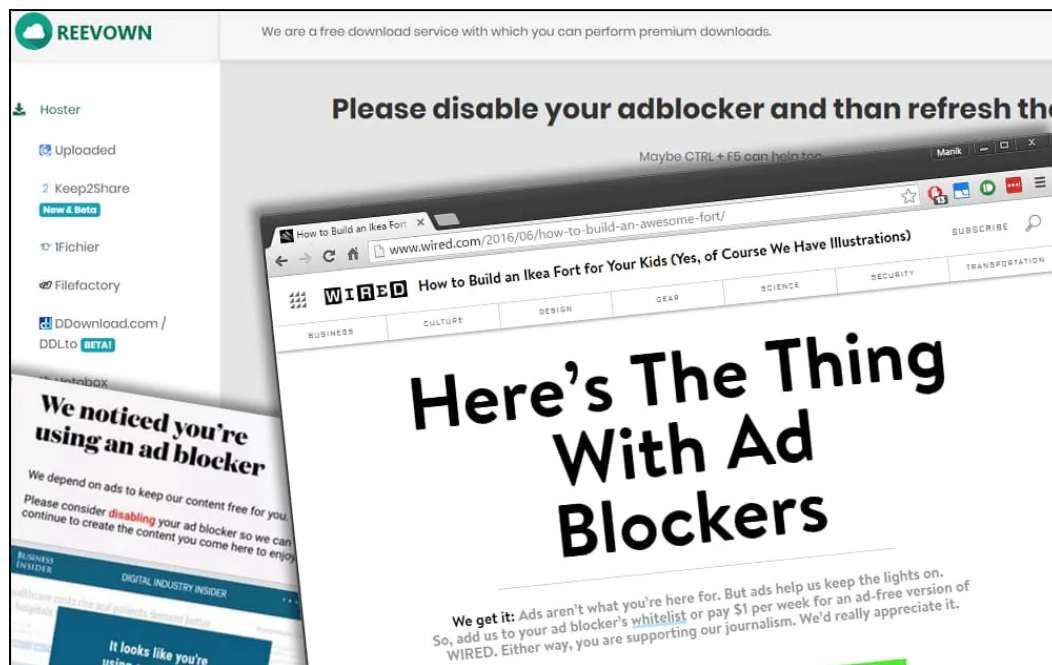


Fighting Adblock-Blocking With Machine Learning

Originally published at <https://www.unite.ai/fighting-adblock-blocking-with-machine-learning/>



A new research initiative from the US and Pakistan has developed a machine learning-based method for identifying websites that are resistant to adblocking and other privacy-preserving technologies, as well as deconstructing the techniques such sites use to 'blend' the origins of ads and real content, so that content is not viewable if ads are blocked.

New adblocking technologies developed from the findings could put an end to incidents where the central content of an article is not viewable when ads are blocked, providing an automated method to separate ad and script resources, rather than the manual approach currently used by popular adblocking frameworks.

The authors conducted a large-scale study of 'mixed resources' on 100,000 websites, finding that 17% of domains, 48% of host-names, 6% of scripts and 9% of content delivery methods deliberately blend tracking (i.e. advertising) functionality with processes that deliver real content. In such cases, article content will disappear for users that are employing adblocking or anti-tracking software, forcing the user to turn off these measures in order to view the content.

In most cases this does not just mean that ads will be visible again, but also that users will be forced back into the cross-domain tracking systems that have [inflamed privacy campaigners](#) in recent years.

The new research offers a system that is able to separate out the components of these 'mixed' web resources at 98% accuracy, allowing adblocking and anti-tracking solutions a chance to disentangle the streams in later iterations of their software, and once again enable content access on adblocked pages.

The [new paper](#) is titled *TrackerSift: Untangling Mixed Tracking and Functional Web Resources*, and comes from researchers at Virginia Tech and UoC Davis in the US, and FAST NUCES and the Lahore University of Management Sciences (LUMS) in Pakistan.

The Adblock Wars

Adblocking systems rely in general on the need for advertising content in a web-page to originate from specific, dedicated domains – generally adtech platforms with domain names and/or IP addresses that can be classed as ‘third party advertising’, allowing the development of blocklists that will not render content from those origins inside a web page.

Additionally, the names of ad-specific resources, such as scripts, can be added to blocklists so that these will not run even in cases where their origins have been deliberately obscured. The naming schemas of such systematically generated scripts are often consistent, enabling recognition and blocklisting.

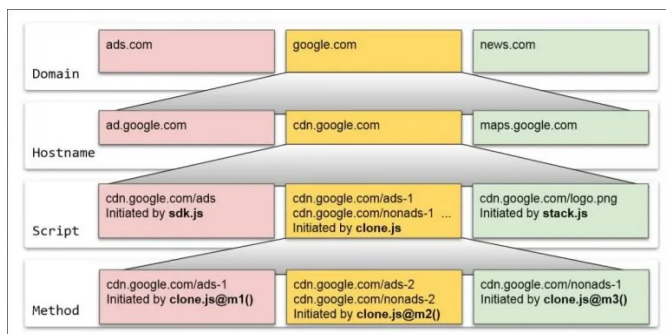
Since an advertisement featured in a web page is frequently chosen in the last few milliseconds of a page load via [dynamic auction processes](#) (based on keywords found in the page, campaign target metrics and many other factors), it’s not practicable to store ads on the host domain, which would in theory impede adblockers from hiding commercial content.

Increasingly, websites are fighting back against adblocking through [CNAME Cloaking](#) – the use of subdomains of the ‘authentic’ domain as proxies to ad servers (i.e., content.example.com will serve ads to example.com, even though the subdomain has no other purpose than to serve advertisements, and is not maintained by the host website, but rather by its advertisers).

However, this method can be quantified and blocked by distinguishing the subdomain’s content as advertising, or using network analysis techniques to identify the subdomain’s anomalous and irregular relationship to the core domain.

TrackerSift

The authors’ paper proposes TrackerSift, a platform to analyze network resources fetched by websites, and then re-categorize mixed resources into ‘content’ and ‘advertising’. At the most general analysis level, TrackerSift records basic network requests for resources, such as ad-content fetched from a Content Delivery Network (CDN) or an advertising platform; but it then drills down to the content of fetched resources, performing code-level analysis, and distinguishing the functions of various types of code calls and procedures.



TrackerSift's analysis hierarchy, from tracking resources (red) through to necessary functional resources (green). Mixed resources, which are likely to lead to content obfuscation (yellow) are subjected to deeper analysis. Source: <https://arxiv.org/pdf/2108.13923.pdf>

Data

To obtain the dataset powering TrackerSift, the authors trawled 100,000 randomly-chosen websites from the 2018 [Tranco top-million list](#). [Selenium](#) browser automation was used together with Google Chrome to perform the task.

The web-crawling network was based on university sites in North America, comprising a 13-node cluster with 112 cores, 52 terabytes of storage and 823 gigabytes of operative RAM among the entire system.

Each node was based in a Docker container and dedicated to crawling a subset of the 100,000 webpages selected, with programmatic pauses for sustainability, and complete erasure of all cookies and identifiers when loading a new domain, to ensure prior sessions and states did not influence the readability of the next domain.

Mixed Scripts

The results show extensive use of *script bundling*, where ad platforms and content hosts deliberately concatenate content-based and ad-based scripts into 'uberscripts' which will impede content display if blocked. By example, the authors note that pressl.co serves a web script bundled via the [WebPack](#) JavaScript concatenator platform, which contains a Facebook tracking pixel, and also code that enables rendering of actual content.

Additionally, the paper notes that a number of domains are willing to embed scripts directly into the code of web pages, making it necessary for adblocking frameworks to address the functionality within the scripts, rather than simply preventing the script from loading based on its third-party source URL.

By localizing these methods, the path is clear for systematic splitting of such code into content and ad categories, and the potential restoration of content display in adblocked environments.

Though existing adblocking solutions, such as NoScript, AdGuard, uBlock Origin and Firefox Smartblock use surrogate scripts which disassemble such merged scripts into blockable component scripts, these depend on [manual rewriting](#) of scripts, leading to an ongoing cold war between the blockers and the ever-shifting techniques that break them. By contrast, TrackerSift offers a potential programmatic method for mixed-content decomposition.

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



Discover More