

ADGRAPH: A Machine Learning Approach to Automatic and Effective Adblocking

Umar Iqbal
The University of Iowa

Zubair Shafiq
The University of Iowa

Peter Snyder
Brave Software

Shitong Zhu
University of California, Riverside

Zhiyun Qian
University of California, Riverside

Benjamin Livshits
Brave Software
Imperial College London

ABSTRACT

Filter lists are widely deployed by adblockers to block ads and other forms of undesirable content in web browsers. However, these filter lists are manually curated based on informal crowdsourced feedback, which brings with it a significant number of maintenance challenges. To address these challenges, we propose a machine learning approach for automatic and effective adblocking called ADGRAPH. Our approach relies on information obtained from multiple layers of the web stack (HTML, HTTP, and JavaScript) to train a machine learning classifier to block ads and trackers. Our evaluation on Alexa top-10K websites shows that ADGRAPH automatically and effectively blocks ads and trackers with 97.7% accuracy. Our manual analysis shows that ADGRAPH has better recall than filter lists, it blocks 16% more ads and trackers with 65% accuracy. We also show that ADGRAPH is fairly robust against adversarial obfuscation by publishers and advertisers that bypass filter lists.

1 INTRODUCTION

Background. Adblocking deployment has been steadily increasing over the last several years. According to PageFair, adblockers are used on more than 600 million devices globally as of December 2016 [15, 24, 26]. There are several reasons that have led to many users installing adblockers like Adblock Plus and uBlock Origin. First, many websites show flashy and intrusive ads that degrade user experience [25, 27]. Second, online advertising has been repeatedly abused by hackers to serve malware (so-called *malvertising*) [44, 48, 54] and more recently *cryptojacking*, where attackers hijack computers to mine cryptocurrencies [37]. Third, online behavioral advertising has incentivized a nefarious ecosystem of online trackers and data brokers that infringes on user privacy [28, 36]. Most adblockers not only block ads, but also block associated malware and trackers. Thus, in addition to instantly improving user experience, adblocking is a valuable security and privacy enhancing technology that protects users against these threats.

Motivation. Perhaps unsurprisingly, online publishers and advertisers have started to retaliate against adblockers. First, many publishers deploy anti-adblockers, which detect adblockers and force users to disable their adblockers. This arms race between adblockers and anti-adblockers has been extensively studied [45, 46]. Researchers have proposed approaches to detect anti-adblocking

scripts [40, 49, 60], which can then be blocked by adblockers. Second, some advertisers have started to manipulate the delivery of their ads to bypass filter lists used by adblockers. For example, Facebook recently obfuscated signatures of ad elements that were used by filter lists to block ads. Adblockers, in response, quickly identified new signatures to block Facebook ads. This prompted a few back and forth actions, with Facebook changing their website to remove ad signatures, and adblockers responding with new signatures [52].

Limitations of Filter Lists. While adblockers are able to block Facebook ads (for now), Facebook’s whack-a-mole strategy points to two fundamental limitations of adblockers. First, adblockers use manually curated filter lists to block ads and trackers based on informally crowdsourced feedback from the adblocking community. This manual process of filter list maintenance is inherently slow and error-prone. When new websites are created, or existing websites make changes, it takes adblocking community some time to catch up by updating the filter lists [1]. This is similar to other areas of system security, such as updating anti-virus signatures [31, 42]. Second, rules defined in these filter lists are fairly simple HTTP and HTML signatures that are easy to defeat for financially motivated publishers and advertisers. Researchers have shown that randomization techniques, where publishers constantly mutate their web content (e.g., URLs, element ID, style class name), can easily defeat signatures used by adblockers [51]. Thus, publishers can easily, and continuously, engage in the back and forth with adblockers and bypass their filtering rules. It would be prohibitively challenging for the adblocking community to keep up in such an adversarial environment at scale. In fact, a few third-party ad “unblocking” services claim to serve unblockable ads using the aforementioned obfuscation techniques [14, 32, 58].

Proposed Approach. In this paper, we aim to address these challenges by developing an automatic and effective adblocking approach. We propose ADGRAPH, which alleviates the need for manual filter list curation by using machine learning to automatically identify effective (both accurate and robust) patterns in the page load process to block ads and trackers. Our key idea is to construct a multi-layer graph representation that incorporates fine-grained HTTP, HTML, and JavaScript information during page load. Combining information across the different layers of the web stack allows us to capture tell-tale signs of ads and trackers. We extract a variety of structural (degree and connectivity) and content (domain and keyword) features from the constructed graph to train a supervised machine learning model to detect ads and trackers.

Technical Challenges. In order to achieve automatic and effective adblocking, ADGRAPH needs to handle two main technical challenges. First, we need to capture fine-grained interactions between DOM elements for tracing the relationships between ads/trackers and the rest of the page content. Specifically, we need to record changes to the DOM, especially in relation to JavaScript code loading and execution. To this end, we use an instrumented version of Chromium web browser to extract HTML, HTTP, and JavaScript information during page load in an efficient manner [43]. Second, we need a reliable ground truth for ads and trackers to train an accurate machine learning model. As discussed earlier, the adblocking community relies on filter lists, which are manually curated based on informal crowdsourced feedback. While the collective amount of work put in by volunteers for maintaining these filter lists is impressive [9, 20], filter lists routinely suffer from errors, which include both false negatives: missed ads *and* false positives: site breakage [1]. To this end, we extract a variety of structural and content features to train a supervised machine learning classification algorithm called Random Forest, which enables us to avoid over-fitting on noisy ground truth [59].

Contributions. We summarize our key contributions as follows.

- (1) We leverage information across multiple layers of the web stack—HTML, HTTP, and JavaScript—to build a fine-grained graph representation. We then extract a variety of structural and content features from this graph representation so that we can use off-the-shelf supervised machine learning algorithms to automatically and effectively block ads and trackers.
- (2) We employ ADGRAPH on Alexa top-10K websites. We show that ADGRAPH successfully replicates the behavior of existing crowdsourced filter lists. We evaluate the contribution of different feature sets that capture HTTP, HTML, and JavaScript information. We show that jointly using these feature sets helps in achieving 97.7% accuracy, 83.0% precision, and 81.9% recall. The accuracy of ADGRAPH also compares favorably to prior machine learning based approaches [30], which reported an accuracy between 57.5% and 81.8%.
- (3) Our manual analysis reveals that more than 65% of ADGRAPH’s reported false positives are in fact false negatives in the crowdsourced filter lists. We further show that about 74% of the remaining false positives are harmless in that they do not cause any visible site breakage.
- (4) We evaluate the effectiveness of ADGRAPH in a variety of adversarial scenarios. Specifically, we test ADGRAPH against an adversary who can obfuscate HTTP URLs (domains and query strings) and HTML elements (ID and class based CSS selectors) [51]. While crowdsourced filter lists are completely bypassed by such adversarial obfuscation, we show that ADGRAPH is fairly robust against obfuscation attempts—ADGRAPH’s precision and recall decrease by at most 7% and 19% for the most aggressive obfuscation, respectively.

Paper organization. The rest of the paper is organized as follows. Section 2 provides some background on adblockers while highlighting some of the key challenges. Section 3 describes our approach to graph construction and featurization, in detail. Section 4 presents experimental evaluation. Section 5 summarizes related work before we conclude in Section 6.

2 BACKGROUND

Adblocking browsers (e.g. Brave) and adblocking extensions (e.g. Adblock Plus and uBlock Origin) use crowdsourced filter lists to block ads and trackers. Examples of popular crowdsourced filter lists include EasyList [6] to block ads, EasyPrivacy [10] to block trackers, and Anti-Adblock Killer [3] to block anti-adblockers. Filter lists contain HTTP and HTML signatures to block ads and trackers. The crowdsourced, signature-based filter list approach to block ads and trackers has significant benefits. Crowdsourced filter lists are updated by volunteers at an impressive pace [9, 20] to cover new ads and trackers as they are encountered and reported by users [1, 7]. Further, the signatures used by these lists are relatively easy for contributors to understand, broadening the number of possible contributors. However, this crowdsourced, signature-based filter list approach entails significant, and possibly intractable, downsides that suggest the need for new adblocking approaches going forward. The remainder of this section details some of the weaknesses in the current filter list approach, while the following section describes our proposed solution.

Bloat. A growing problem with crowdsourced, signature-based filter lists is that they bloat over time, becoming more difficult to maintain. This is partially a result of the non-expert, non-methodical way of adding signatures to the list. Rules are added far more frequently than rules are removed, resulting in a long tail of stale, low-utility rules. During a crawl of the Alexa top-1K websites, where we visited the home page and three randomly selected child pages, we observed that only 884 of the 32,218 (less than 3%) HTTP rules in EasyList trigger. Additionally, because these increasingly bloated filter lists are manually maintained, they are slow to catch up when websites introduce changes. Iqbal et al. [40] reported that filter lists sometimes take several months to add rules for blocking new anti-adblockers.

Accuracy. Crowdsourced filter lists also suffer in terms of accuracy (both precision and recall). Precision suffers from overly broad rules. This is evidenced by the ever-growing number of exception rules in these filter lists, which are added to undo incorrectly blocked resources from other overly broad rules. As one notable example, 263 exception rules in EasyList exist only to allow resources blocked by other rules *for a single domain*. EasyList contains many such examples of exception rules catering for other overly broad rules. Recall suffers mainly due to the lack of sufficient feedback for less popular websites. Englehardt and Narayanan [36] reported that filter lists are less effective at blocking obscure trackers.

Evasion. Crowdsourced filter lists can be easily evaded by publishers and advertisers. Simple randomization of HTTP and HTML information can successfully evade filter lists used by adblockers [51]. Some publishers have already begun using these techniques. For example, Facebook recently manipulated HTML element identifiers that were being used by EasyList to get their ads past adblockers [33, 47, 52]. Domain generation algorithm (DGA) is another technique that allows publishers and advertisers to evade filter lists [35, 58]. Third-party services such as Instart Logic provide ad “unblocking” services [14] that leverage similar HTTP and HTML obfuscation techniques to evade filter lists.

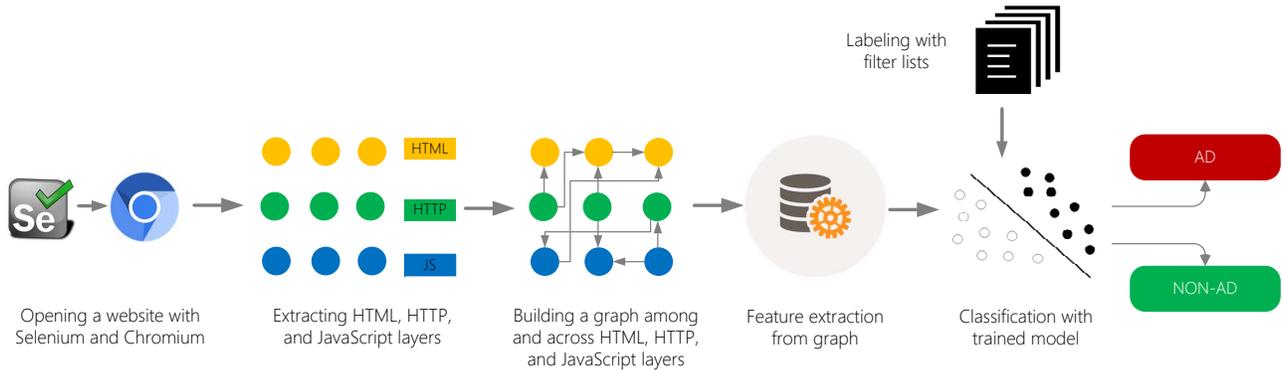


Figure 1: Our proposed approach for classification of AD and NON-AD URLs. We combine information from HTML, HTTP, and JavaScript layers and represent connections among and across these layers in form of a graph. We then extract features from the graph and train a machine learning model to classify AD and NON-AD URLs.

In summary, filter lists used by adblockers are inefficient, unreliable, and susceptible to obfuscation by financially motivated adversaries. We believe that adblocking is a valuable security and privacy-enhancing technology that needs to move beyond crowd-sourced filter lists to stay viable as the arms race between advertisers and adblockers continues to escalate.

3 PROPOSED APPROACH: ADGRAPH

After providing an overview of AdGRAPH’s architecture, we discuss the construction of multi-layer graph representation that incorporates fine-grained HTTP, HTML, and JavaScript information during page load. We then discuss AdGRAPH’s featurization approach to train a machine learning model for detecting ads and trackers.

3.1 Overview

Figure 1 shows an architectural overview of AdGRAPH—a graph-based machine learning approach for detecting ads and trackers on a given web page. AdGRAPH uses an instrumented version of Chromium web browser to extract HTML, HTTP, and JavaScript information during page load. The HTML, HTTP, and JavaScript *layers* of the web stack are processed and converted into a graph structure. This graph structure captures both relationships *within* the same layer (e.g. parent-child HTML elements) and *across* layers (e.g. which JavaScript code units created which HTML elements) of the web stack. We compute a variety of structural (degree and connectivity) and content (domain and keyword) features, and use these features to classify nodes as either AD or NON-AD using a supervised machine learning model.

3.2 HTML, HTTP, and JavaScript Layers

We use an instrumented version of Chromium web browser to build a graph of the relationships between HTML, HTTP, and JavaScript actions related to a web page [43].

HTML. For HTML layer extraction, we inject and execute a script in the website and traverse the complete DOM tree to extract all HTML elements. We start from the root of the DOM tree and traverse it in the breadth-first order. At each level of the DOM tree, we record `outerHTML` (the serialized HTML fragment, which includes its complete contents) and `baseURI` (to link elements to where they are loaded from) attributes of all HTML elements.

JavaScript. For JavaScript layer extraction, we leverage the logs obtained from the instrumented version of Chromium. Specifically, the instrumented Chromium browser records all interactions between JavaScript snippets and HTML elements. The instrumentation extends Chromium’s DevTools to log the addition of new HTML elements, modification to existing HTML elements, and event listeners attached to HTML elements via JavaScript. From these logs, we extract JavaScript snippets, URLs that load the JavaScript snippets, and interactions between JavaScript snippets and HTML elements.

HTTP. For HTTP layer extraction, we rely on information already obtained during extraction of HTML and JavaScript layers. More specifically, we use `baseURI` attribute extracted from the HTML layer and HTTP URLs extracted from the JavaScript layer to construct HTTP layer. We further enrich the HTTP layer by adding HTTP URLs from `href` and `src` tags of HTML elements.

3.3 Graph Construction

We next combine information from HTML, HTTP, and JavaScript layers. To this end, we represent information from different layers in the form of a graph structure. As we elaborate below, each HTML element, HTTP URL, and JavaScript snippet is represented as a node, and relationships among these nodes are represented as edges. By way of illustration, Figure 2 visualizes the graph for a toy example in Code 1.

3.3.1 Graph Nodes. We categorize nodes in the graph according to their layer information. We have three types of nodes: HTML

```

1 <html>
2 <head>
3   <script src="thirdparty.com/script1.js" type="text/
4     javascript"></script>
5   <script src="thirdparty1.com/script2.js" type="text/
6     javascript"></script>
7   <script type="text/javascript">
8     var iframe=document.createElement('iframe');
9     var html='';
10    iframe.src='adnetwork.com';
11    document.body.appendChild(iframe);
12  </script>
13  <link rel="stylesheet" href="../style1.css">
14 </head>
15 <body>
16   <div class = "class1" id = "id1">
17     
19     ...
20   </div>
21   <div class = "class1" id = "id2">
22     <iframe id = "iframe1" src = "adnetwork.com">
23     ...
24   </iframe>
25   <div class = "class2" id = "id3">
26     <button type="button" onclick="func()"></button>
27     ...
28     <ul>
29       <li>List item.</li>
30     </ul>
31   </div>
32 </div>
33 ...
34 </body>
35 </html>

```

Code 1: A sample HTML page.

element, HTTP URL, and JavaScript snippet. For the example graph in Figure 2, HTML nodes are represented by ●, HTTP nodes are represented by ●, and JavaScript nodes are represented by ●. We further categorize HTML, HTTP, and JavaScript nodes to capture more fine-grained information about them.

HTML element nodes. HTML element nodes are further categorized as HTML iframe element, HTML image element, HTML style element, and HTML miscellaneous element based on their tag name information. HTML iframe element nodes are HTML element nodes with iframe tag name. HTML image element nodes are HTML element nodes with img tag name. HTML style element nodes are HTML element nodes with style tag name. HTML miscellaneous element nodes are HTML element nodes with tag name other than iframe, img, or style.

HTTP URL nodes. HTTP URL nodes are further categorized according to their edges with HTML element nodes and JavaScript snippet nodes. We categorize HTTP URL nodes into HTTP script URL, HTTP source URL, HTTP iframe URL, and HTTP element URL nodes. HTTP script URL nodes are HTTP URL nodes that load JavaScript snippets. HTTP source URL nodes are HTTP URL nodes that load all categories of HTML elements other than iframe HTML elements. HTTP iframe URL nodes are HTTP URL nodes that load iframe HTML elements. HTTP element URL nodes are HTTP URL nodes that are present as src of HTML elements.

JavaScript snippet nodes. JavaScript snippet nodes are further categorized according to their scope: inline or referenced. JavaScript inline snippet nodes are JavaScript snippet nodes that are present

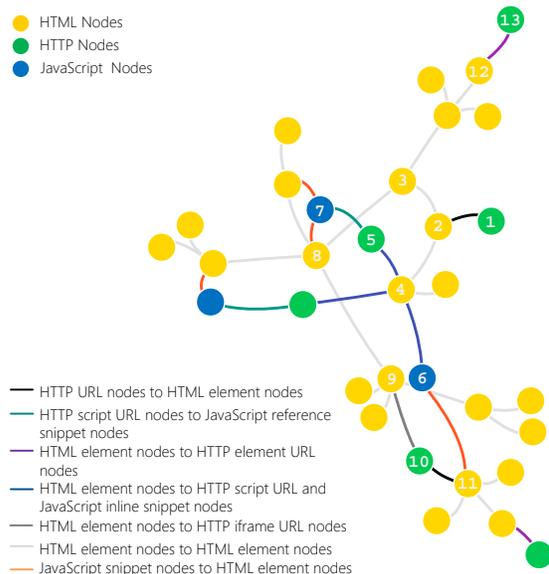


Figure 2: The three-layered graph for Listing 1. The graph shows HTML, HTTP, and JavaScript layers, along with edges, both within and across these layers.

as inline scripts in the DOM. JavaScript reference snippet nodes are JavaScript snippet nodes that are loaded as reference in the DOM.

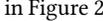
3.3.2 Graph Edges. We represent relationships among nodes within each layer and across different layers as edges. We categorize edges based on the relationships between pairs of nodes into seven broad categories, which further have subcategories. Figure 2 shows the categories of edges in the example graph using different colors.

Edges from HTTP URL nodes to HTML element nodes. Edges from HTTP URL nodes to HTML element nodes represent the loading of an HTML element from an HTTP URL. These edges capture information about the origin of HTML elements. These edges are further categorized into edges from HTTP iframe URL nodes to HTML iframe element nodes and edges from HTTP source URL nodes to all other categories of HTML element nodes. These edges are represented by — in Figure 2, which has two such edges: one from node 1 to node 2 and the other from node 9 to node 10. The edge from node 1 to node 2 represents the initial page load and the edge from node 9 to node 10 represents the loading of an iframe. Node 2 is the base HTML element and it corresponds to the line 1 of Code 1. Node 10 is an HTML iframe element node with id iframe1 and it corresponds to the line 20 of Code 1.

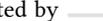
Edges from HTTP script URL nodes to JavaScript reference snippet nodes. Edges from HTTP script URL nodes to JavaScript reference snippet nodes represent the loading of a JavaScript snippet from an HTTP URL. These edges capture information about the

origin of JavaScript snippets. These edges are represented by  in Figure 2. The edge from node 5 to node 7 in Figure 2 represents the loading of a third-party script. Node 5 is an HTTP script URL node and node 7 is a JavaScript reference snippet node. Both node 5 and node 7 correspond to the `script` item on line 3 of Code 1.

Edges from HTML element nodes to HTTP element URL nodes. Edges from HTML element nodes to HTTP element URL nodes represent an HTML element that has an HTTP URL as its source. These edges capture information about the content loaded inside an HTML element. These edges are further categorized into edges from HTML image element nodes to HTTP element URL nodes, edges from HTML style element nodes to HTTP element URL nodes, and edges from HTML miscellaneous element nodes to HTTP element URL nodes. These edges are represented by  in Figure 2. The edge from node 12 to node 13 in Figure 2 represents the loading of an image. Node 12 is an HTML image element node and node 13 is an HTTP element URL node. Both node 12 and node 13 correspond to the `img` item with id `image1` on line 16 of Code 1.

HTML element nodes to HTTP script URL and JavaScript inline snippet nodes. Edges from HTML element nodes to HTTP script URL and JavaScript inline snippet nodes capture the occurrence of JavaScript snippets in the the DOM tree. These edges are represented by  in Figure 2. We have two such edges in Figure 2: one from node 4 to node 5 and the other from node 4 to node 6. The edge from node 4 to node 5 represents the occurrence of a script with third-party reference and the edge from node 4 to node 6 represents the occurrence of an inline script. Node 4 is an HTML element node and it corresponds to the `head HTML` item on line 2 of Code 1. Node 5 is an HTTP script URL node and it corresponds to the `script` item on line 3 of Code 1. Node 6 is a JavaScript inline snippet node and it corresponds to the `inline script` item on line 5 of Code 1.

HTML element nodes to HTTP iframe URL nodes. Edges from HTML element nodes to HTTP iframe URL nodes represent the loading of an HTML iframe element from an HTTP URL. These edges capture information about the origin of HTML iframe elements and are represented by  in Figure 2. The edge from node 9 to node 10 in Figure 2 represents the loading of an iframe. Node 9 is an HTML element node and it corresponds to the `div HTML` item with id `id2` on line 19 of Code 1. Node 10 is an HTTP iframe URL node and it corresponds to the `iframe` item with id `iframe1` on line 20 of Code 1.

Edges from HTML element nodes to HTML element nodes. Edges from HTML element nodes to HTML element nodes capture the hierarchy of HTML elements in the DOM tree. These edges capture the parent-child relationship among parent HTML elements and child HTML elements. These edges are represented by  in Figure 2. A majority of edges in Figure 2 are edges from HTML element nodes to HTML element nodes. One such edge from node 3 to node 8 in Figure 2 represents the parent-child relationship between two HTML elements. Node 3 is an HTML element node and it corresponds to the `body HTML` item on line 14 of Code 1. Node 8 is an HTML element node and it corresponds to the `div HTML` item with id `id2` on line 19 of Code 1.

Edges from JavaScript source snippet nodes to HTML element nodes. Edges from JavaScript source snippet nodes to HTML element nodes represent the interaction between a script and an HTML element. These edges capture addition of new HTML elements, modification to existing HTML elements, and event listeners attached to HTML elements. There can be multiple edges from one JavaScript source snippet node to an HTML element node. These edges are represented by  in Figure 2. The edge from node 6 to node 11 in Figure 2 represents the addition of an HTML iframe element with further sub HTML image element. Node 6 is a JavaScript inline snippet node and it corresponds to the inline script item on line 5 of Code 1. Node 11 is an HTML iframe element node with id `iframe1` and it corresponds to the line 20 of Code 1.

3.4 Feature Extraction

After constructing the graph, we are set to extract features from it to train a machine learning model for classifying ads and trackers. We extract different structural (degree and connectivity) and content (domain and keyword) features for nodes in the graph. Degree features include attributes such as in-degree, out-degree, number of descendants, and number of node additions/modifications. Connectivity features include a variety of centrality metrics. Domain features capture first versus third party domain/sub-domain information. Keyword features capture the presence of certain ad-related keywords in query string. Below, we explain each of these features in detail.

3.4.1 Degree Features. Degree features provide information about the number of edges incident on a node. Below we explain specific degree metrics that we extract as features.

- **In-Degree:** The in-degree of a node is defined as the number of inward edges incident on the node. We separately compute in-degree for different edge types based on the type of node they are originating from. We also use aggregate node in-degree.
- **Out-Degree:** The out-degree of a node is defined as the number of outward edges incident on the node. We separately compute out-degree for different edge types based on the type of their destination node. We also use aggregate node out-degree.
- **Descendants:** The descendants of a node is defined as the number of nodes reachable from it.
- **Addition of nodes:** For a node, we count the number of nodes added by it. This feature specifically captures the addition of new DOM nodes by a script.
- **Modification of node attributes:** For a node, we count the number of node attribute modifications made by it. This feature specifically captures the attribute modifications of DOM nodes by a script. We consider modifications to existing attributes, addition of new attributes, and removal of existing attributes.
- **Event listener attachment:** For a node, we count the number of event listeners attached by it. This feature specifically captures the event listener attachments to DOM nodes by a script.

3.4.2 Connectivity Features. Connectivity features provide information about the relative importance of a node in the graph. Below we explain specific connectivity metrics that we extract as features.

- **Katz centrality:** The Katz centrality of a node is a centrality-based measure of relative importance. It is influenced by the degree of a node and degree of its neighboring nodes and degree of nodes reachable by its neighboring nodes.
- **Closeness centrality:** The closeness centrality of a node is defined as the average length of shortest paths to all nodes reachable from it.
- **Mean degree connectivity:** The mean degree connectivity of a node is defined as the average of degrees of all its neighboring nodes.
- **Eccentricity:** The eccentricity of a node is defined as the maximum of distance from that node to all the other nodes in a graph.

3.4.3 *Domain Features.* Domain features provide information about the domain specific properties of a node’s associated URLs. Below we explain specific domain properties that we extract as features.

- **Domain party:** For a node, the domain party feature describes whether the domain of the node URL is first-party or third-party.
- **Sub-domain:** For a node, the sub-domain feature describes whether the domain of the node URL is a sub-domain of the first-party domain.
- **Base domain in query string:** For a node, the base domain in query string feature describes whether the node URL has base domain as a parameter in query string.
- **Same base domain and request domain:** For a node, the same base domain and request domain feature describes whether the domain of node URL is same as the base domain.
- **Node category:** For a node, the node category feature describes the HTTP node type of the node URL as described in Section 3.3.1.

3.4.4 *Keyword Features.* Keyword features provide information about the use of certain keywords in URLs. Below we explain specific keyword patterns that we extract as features.

- **Ad keywords:** For a node, we capture the number of ad-related keywords present in the node URL. We use keywords such as ‘advertise’, ‘banner’, and ‘advert’ because they tend to frequently appear in advertising URLs. We also capture the number of ad-related keywords that are followed by a special character (e.g., ‘;’ and ‘=’) in the node URL. This helps us exclude scenarios where ad-related keywords are part of non ad-related text in the URL.
- **Query string parameters:** For a node, we count the number of semicolon separated query string parameters in the node URL. We also capture whether query string parameters are followed by a ‘?’ and they are separated by a ‘&’.
- **Ad dimension information in query string:** For a node, we check whether the query string has ad dimensions. We define a pattern of 2–4 numeric digits followed by the character x and then again followed by 2–4 numeric digits as the presence of ad size in a query string parameter. We also capture the presence of screen dimension in a query string parameter. We look for keywords such as screenheight, screenwidth, and screendensity.

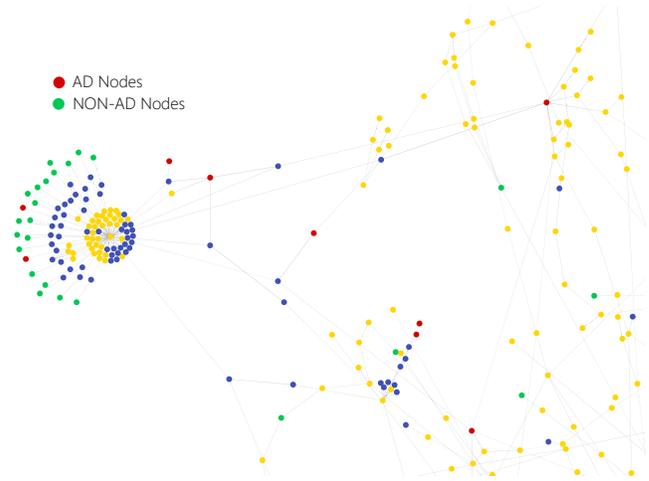


Figure 3: Zoomed-in version of the graph constructed for www.bbc.com.

3.5 Feature Analysis

Next we analyze a few of the features that we use to train our machine learning classifier. Consider the graph shown in Figure 3 as a reference example during our feature analysis. While nodes in Figure 3 follow the color scheme explained in Section 3.3.1, ad/tracker nodes (AD) are represented by ● and non ad/tracker nodes (NON-AD) are represented by ●. We explain the ground truth labeling of AD and NON-AD nodes further in the next section.

Figure 4(a) plots the cumulative distribution function (CDF) of closeness centrality. We note that AD nodes tend to have higher closeness centrality values as compared to NON-AD nodes. It means that AD nodes are more well connected than NON-AD nodes. In Figure 3, AD nodes are generally connected to multiple HTML element nodes and JavaScript snippet nodes as compared to NON-AD nodes, which mostly appear as leaf nodes. These additional connections enable extra paths for AD nodes, making them more central than NON-AD nodes.

Figure 4(b) plots the CDF of eccentricity of a node. As shown in Figure 3, AD nodes are mostly accompanied by JavaScript snippet nodes that represent analytics scripts. Thus, AD nodes have more paths to reach other nodes in the graph. Analytics scripts usually appear at the start of DOM tree and AD nodes connected to them will have shorter paths (low eccentricity) to other nodes compared to NON-AD nodes that appear as leaf nodes.

Figure 4(c) plots the CDF of number of descendants of a node. The number of descendants of a node provides a clear distinction between AD and NON-AD nodes. As we note in Figure 3, most NON-AD nodes appear as leaf nodes (image URLs, anchor URLs) while most AD nodes appear as non-leaf nodes. This behavior is captured by the number of descendants of a node.

Figure 4(d) plots the distribution of first-party versus third-party URLs for AD and NON-AD nodes. As expected, it is evident that most ads and trackers are loaded by third-party URLs.

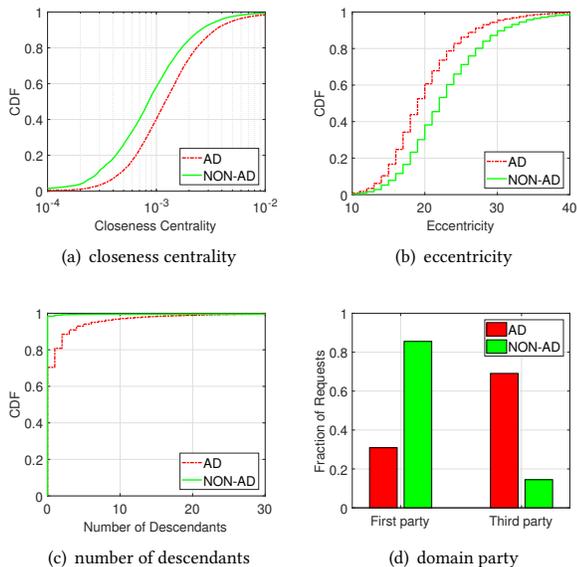


Figure 4: Conditional feature distributions.

3.6 Supervised Classification

We use random forest [34] which is a well-known ensemble supervised learning algorithm for classification. Random forest combines decisions from multiple decision trees, each of which is constructed using a different bootstrap sample of the data, by choosing the mode of the predicted class distribution. Each node for a decision tree is split using the best among the subset of features selected at random. This feature selection mechanism is known to provide robustness against over-fitting issues. We configure random forest as an ensemble of 10 decision trees with each decision tree trained using $\text{int}(\log M + 1)$ features, where M is the total number of features.

4 EVALUATION

We first evaluate AdGRAPH by measuring how closely its classification matches popular crowdsourced filter lists in blocking ads and trackers. AdGRAPH replicates the behavior of crowdsourced filter lists with 97.7% accuracy, 83.0% precision, and 81.9% recall. We next manually analyze disagreements between classifications by AdGRAPH and crowdsourced filter lists. We find that more than 65% of AdGRAPH’s false positives are in fact false negatives of crowdsourced filter lists. We also evaluate the effectiveness of AdGRAPH against adversarial obfuscation that attempts to bypass crowdsourced filter lists. We show that AdGRAPH is fairly robust against obfuscation attempts, the precision decreases by 7% while the recall decreases by at most 19% for the most aggressive obfuscation. Overall, our experimental evaluation shows that AdGRAPH’s graph-based machine learning approach outperforms manually curated, crowdsourced filter lists in terms of both accuracy and robustness.

4.1 Experimental Setup

Ground Truth. We use an instrumented version of Chromium web browser [43] with Selenium WebDriver to automatically crawl the home pages of Alexa top-10K websites. We are able to construct graphs incorporating information across HTML, HTTP and JavaScript layers of the web stack for 7,699 websites.¹ We need a web-scale “ground truth” to evaluate the accuracy of AdGRAPH in blocking ads and trackers. We use the union of 9 popular crowdsourced filter lists² despite their well-known shortcomings for two reasons. First, the popularity of these crowdsourced filter lists suggests that they are reasonably accurate even though they are imperfect. Second, a better alternative—building a web-scale, expert-labeled ground truth—would require funding and labor at a scale not available to the research community. Thus, despite their shortcomings, we make the methodological choice to treat these crowdsourced filter lists as ground truth for labeling ads and trackers. Using these crowdsourced filter lists, we label HTTP URL nodes as AD or NON-AD. While we do not label non-HTTP nodes, we do use HTML and JavaScript layers to enhance the constructed graph by capturing fine-grained information flows across HTTP, HTML, and JavaScript layers. Overall, our ground truth labeled data set has 131,917 AD and 1,906,763 NON-AD nodes.

Cross-Validation. We train a machine learning model to detect AD and NON-AD nodes. Specifically, we train a single model using the available labeled data that would generalize to any unseen website. We train and test AdGRAPH using stratified 10-fold cross-validation. We use 9 folds for training AdGRAPH’s machine learning model and the leftover fold for testing it on unseen websites. We repeat this process 10 times using different training and test folds.

Accuracy Results. We measure the accuracy of AdGRAPH in terms of true positive (TP), false negative (FN), true negative (TN), and false positive (FP).

- **TP:** An ad or tracker is correctly labeled AD.
- **FN:** An ad or tracker is incorrectly labeled NON-AD.
- **TN:** A non-ad or non-tracker is correctly labeled NON-AD.
- **FP:** A non-ad or non-tracker is incorrectly labeled AD.

Overall, AdGRAPH achieves 97.7% accuracy, with 83.0% precision and 81.9% recall. Figure 5 shows the Receiver Operating Characteristic (ROC) curve for AdGRAPH. The curve depicts the trade-off between true positive rate and false positive rate as the discrimination threshold of our classifier is varied. Using this formulation, we can also quantify the accuracy of AdGRAPH by using the area under the ROC curve (AUC). High AUC values reflect high positive rate and low false positive rate. AdGRAPH’s true positive rate quickly converges to near 1.0 for fairly small false positive rates with AUC = 0.98.

¹Note that we are unable to crawl 1,275 websites due to server-side errors (e.g., HTTP error codes 404 and 500) and 1,026 websites due to issues in the instrumented browser which is meant to be a research prototype. Our eyeball analysis showed that there was not something specific in the nature of functionality of these sites that would bias our evaluation.

²We use EasyList [6], EasyPrivacy [10], Anti-Adblock Killer [2], Warning Removal List [21], Blockzilla [4], Fanboy Annoyances List [11], Fanboy Social List [12], Peter Lowe’s list [16], and Squid Blacklist [17].

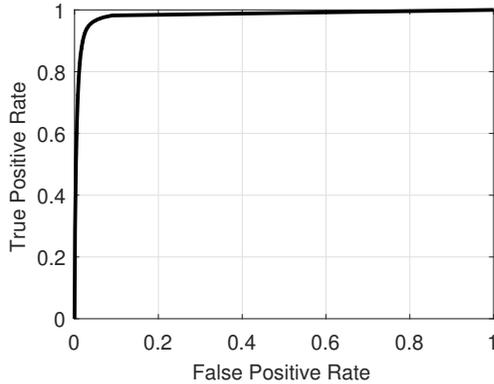


Figure 5: ROC curve of AdGRAPH for detecting ads and trackers.

Ablation Results. In order to maximize the accuracy of AdGRAPH, we combine four categories of features (degree, connectivity, domain, and keyword) defined in Section 3.4. We now evaluate the relative importance of different feature categories. To this end, we compare the precision and recall of all possible combinations of different feature categories. Specifically, we first use each feature category individually, then in combinations of two and three, and finally use all of them together. Figure 6 reports the precision and recall of different feature category combinations. For individual feature categories, none of them provide sufficient accuracy to be used stand-alone with the exception of connectivity features. We note that keyword features (also used in prior work [30]) have the highest precision but lowest recall as compared to other feature categories. As we start combining features, we find that precision and recall improve by the addition of each feature category, and that precision and recall are maximized by combining all four feature categories.

4.2 False Positive Analysis

When AdGRAPH detects NON-AD nodes (labeled by crowdsourced filter lists) as AD nodes, the disagreement is recorded as a false positive. Due to the previously discussed shortcomings of crowdsourced filter lists as ground truth, we manually evaluate a sample of these false positives. Our analysis reveals that in most cases AdGRAPH makes the right evaluation, and that the filter lists are incorrect—we refer to these cases as “false, false-positives.” Overall, we find that 65.4% of AdGRAPH’s false positives are in fact “false, false-positives.” Therefore, AdGRAPH’s precision reported in Section 4.1 is a lower bound on actual precision. Next, we discuss our methodology to analyze AdGRAPH’s false positives with a few illustrative examples.

For 22,062 cases, AdGRAPH has a disagreement with the ground truth label by crowdsourced filter lists. To analyze these disagreements, we group 22,062 URLs into 3,950 clusters according to their base domain because resources from the same base domain are likely to provide similar functionality. For example, `evil.com/script1.js` and `evil.com/script2.js` are more likely to have a similar functionality as compared to `nice.com/script1.js`. We then manually analyze 1,400 of

these 3,950 groups, which comprise of 11,714 unique URLs. Due to a large number of these URLs, we only sample one URL out of each cluster for manual analysis.

4.2.1 Methodology. We manually analyze the disagreements between AdGRAPH and the filter lists as follows.

- (1) If the URL contains keywords associated with advertising (e.g. an ad exchange) or tracking (e.g. analytics), we consider AdGRAPH’s labeling correct.
- (2) If we find that the URL is present in less popular regional filter lists [8] or it is mentioned on adblocking forums [1], we consider AdGRAPH’s labeling correct.
- (3) If we find ad and tracking related keyword in JavaScript served by URLs, we consider AdGRAPH’s labeling correct.
- (4) Otherwise, we consider AdGRAPH’s labeling incorrect.

4.2.2 Results. Table 1 shows the breakdown of our manual analysis. ‘N/A’ refers to URLs that are inaccessible due to server downtime. ‘Unknown’ refers to URLs that are difficult to analyze due to code obfuscation. ‘Functional’ refers to URLs that are not associated with ads or trackers. Interestingly, 65.4% of the sampled false positive cases are confirmed to be “false, false-positives.” Thus, AdGRAPH is able to automatically leverage structural information through machine learning to identify many ads and trackers that are otherwise missed by crowdsourced filter lists.

4.2.3 Case Studies. Below we discuss a few interesting examples of “false, false-positives.” We will discuss AdGRAPH’s actual false positives later in Section 4.3.

yimg.com. We first discuss a case of conversion tracking script in Code 2 from yimg.com that is detected by AdGRAPH but missed by crowdsourced filter lists. Besides code analysis, yimg also confirms this tracking module in its official documentation [22]. We find 42 different websites in our sample that load this script. Given the popularity of this unobfuscated tracking script, it is surprising that none of the filter lists block it. This case study further highlights the benefit of AdGRAPH’s automated detection of ads and trackers.

digitru.st. Next, we discuss a script in Code 3 that claims to provide “anonymous” tracking service to publishers [5]. Beyond tracking, we also identify anti-adblocking functionality in the script. We find 10 different websites in our sample that load this script. However, it is not blocked by any of the filter lists.

glbimg.com. Next, we discuss a tracking script in Code 4 from glbimg.com. The script sends the recorded page views to the server. We find 5 different websites in our sample that load this script.

intellicast.com. Next, we discuss an ad loading script in Code 5 from intellicast.com. The script serves as an “entry point” to load ads on a web page. We visually compare the web pages with and without the script, and confirm that we are able to block ads that are loaded by the script. The visual comparison can be seen in Figure 7.

Functional Advertising/Tracking	N/A	Unknown
427 (30.5%)	915 (65.4%)	23 (1.6%) 35 (2.5%)

Table 1: Breakdown of false positive analysis

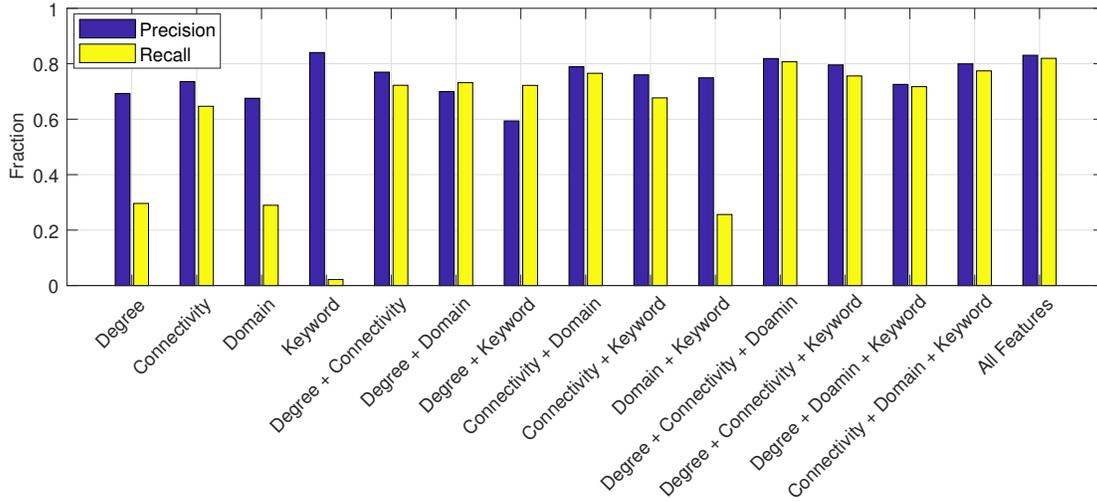


Figure 6: Precision and recall using different feature categories.

```

1 function e(e, a) {
2   if (e.google_conversion_language =
3     window.yahoo_conversion_language,
4     e.google_conversion_color =
5     window.yahoo_conversion_color,
6     e.google_conversion_label =
7     window.yahoo_conversion_label,
8     e.google_conversion_value =
9     window.yahoo_conversion_value,
10    e.google_conversion_domain = a,
11    e.google_remarketing_only = !1, "function" ==
12    typeof window.google_trackConversion)
13     window.google_trackConversion(e);
14   else {
15     var i = o(a, "conversion_async.js");
16     n(i, function() {
17       "function" == typeof
18         window.google_trackConversion &&
19         window.google_trackConversion(e)
20     })
21   }
22 }

```

Code 2: yimg.com ads conversion tracker loader script.

```

1 function b(k) {
2   k = a.call(this, k) || this;
3   void 0 === b.instance && (b.instance = k,
4     f.noAutoStartTracker ||
5     k.client.sendPageView(k.makeParams()));
6   return b.instance
7 }

```

Code 4: glbimg.com tracking script.

```

1 createClass(MoneyTreeBase, [{
2   key: 'getSlots',
3   value: function getSlots() {
4     return new Promise$(function (resolve) {
5       return gptCmd(function () {
6         return resolve(window.googletag.pubads
7           ().getSlots() || []);
8       });
9     });
10  }

```

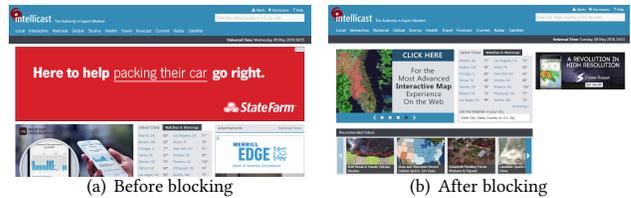
Code 5: intellicast.com ads loader script.

```

1 ad blocker: {
2   detection: !1,
3   blockContent: !1,
4   userMessage: "Did you know advertising pays for
5     this brilliant content? Please disable your
6     ad blocker, then press the Reload button
7     below ... and thank you for your visit!",
8   popupFontColor: "#5F615D",
9   popupBackgroundColor: "#FFFFFF",
10  logoSrc: null,
11  logoText: null,
12  pictureSrc: null
13 },

```

Code 3: digitru.st tracking script.



(a) Before blocking (b) After blocking

Figure 7: Effect of blocking “false, false-positive” on intellicast.com.

4.3 Breakage Analysis

We now analyze the impact of AdGRAPH’s actual false positives on site breakage. To this end, we manually open websites and observe any breakage caused by the blocking of AdGRAPH’s actual

false positives. Specifically, we identify the breakage by analyzing whether the false positives remove resources that are critical for the site’s functionality. We manually analyze 528 websites from 427 clusters, which are classified as ‘Functional’ false positives in

Table 1. Note that we could not analyze 183 websites for breakage because of the use of dynamic URLs. Interestingly, we find that ADGRAPH’s actual false positives do not result in any visible breakage for more than 74% of the websites. Therefore, we conclude that a majority of ADGRAPH’s actual false positives in fact do not harm user experience. Next, we discuss our methodology to analyze the breakage caused by ADGRAPH’s actual false positives.

4.3.1 *Methodology.* We manually analyze the site breakage as follows.

- (1) We first generate a custom filter list to block ADGRAPH’s actual false positives.
- (2) We then open the same website on two browser instances, one with adblocker and the other without adblocker.
- (3) We compare the two loaded web pages, perform 3–5 clicks in different regions and scroll up/down 3 times to trigger any potential breakage.

We infer breakage by looking for visual inconsistencies across the two browser instances. We specify three levels for site breakage: none, minor, and major. These three levels respectively correspond to no visible inconsistencies, minor visible inconsistencies (e.g. few images disappear), and page malfunction (e.g. garbled layout).

4.3.2 *Results.* Table 2 shows the breakdown of our manual analysis of site breakage. Interestingly, we do not observe any breakage for 74.7% of the websites. We do observe minor breakage on 19.1% of the websites and major breakage on 6.1% of the websites. We argue this is a reasonably small number, showing limited impact of ADGRAPH’s actual false positives on user experience. In future, we plan to further reduce such breakage by incorporating additional features.

Not visible	Visible: minor	Visible: major
247 (74.7%)	66 (19.1%)	21 (6.1%)

Table 2: Breakdown of breakage analysis results.

4.3.3 *Case Studies.* Below we discuss a few interesting examples of ADGRAPH’s actual false positives. We discuss cases for different breakage levels and provide some insights into their causes and impact on user experience.

urbandictionary.com. We first discuss ADGRAPH’s actual false positive on urbandictionary.com that does not cause any breakage. urbandictionary.com loads a JavaScript library named twemoji to support emojis [19]. Interestingly, removing this library does not cause any visible breakage on the website.

darty.com. ADGRAPH blocks a functional script on darty.com without causing any visible breakage. Our manual analysis shows that it is a helper module to manage browser cookies for shopping cart functionality. We do not find any functional breakage in multiple test runs.

fool.com. ADGRAPH blocks a script called FontAwesome on fool.com while causing minor breakage. Our manual analysis shows that it helps with display of vector icons and social logos [13]. As

shown in Figure 8, blocking the script only changes icons used on the web page while not affecting any major functionality.

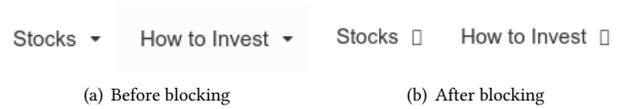


Figure 8: Minor breakage on fool.com.

meteored.mx. ADGRAPH blocks an iframe served by meteored.mx on noticiaaldia.com causing minor breakage. We note that it is a weather widget as shown in Figure 9. Blocking the iframe only causes the weather widget to disappear.



Figure 9: Weather widget iframe from meteored.mx.

game8.jp. ADGRAPH blocks a script called Bootstrap on game8.jp causing major breakage. As shown in Figure 10, blocking Bootstrap garbles the web page and major portion of the web page goes missing.



Figure 10: Major breakage on game8.jp.

4.4 Robustness to Obfuscation

Next, we analyze the robustness of ADGRAPH against adversarial obfuscation by publishers and advertisers to get their ads past adblockers. We focus our attention on practical obfuscation techniques that have been recently developed and deployed to bypass filter lists used by adblockers. Wang et al. [51] recently proposed to

randomize HTTP URLs and HTML element information to this end. Similar obfuscation techniques are being used by ad “unblocking” services [14, 32, 58] to bypass filter lists used by adblockers. We implement these obfuscation attacks on Alexa top-10K websites and evaluate the robustness of AdGRAPH against them.

4.4.1 HTML Element Obfuscation Attacks. First, we implement an adversarial publisher who can manipulate attributes of HTML elements. HTML element hiding rules in filter lists typically use element `id` and `class` attributes. Thus, an adversarial publisher can simply randomize `id` and `class` attributes at the server side to bypass these HTML element hiding rules. It is noteworthy, however, that modifications of `id` and `class` attributes need to be constrained so as to not impact the appearance and functionality of the website. For example, a script that interacts with an HTML element based on its attribute will not be able to interact if HTML element attributes are randomized. To address this issue with obfuscation, Wang et al. [51] demonstrated a workable solution that requires overriding the relevant JavaScript APIs and keeping a map of original attribute values. To evaluate the robustness of AdGRAPH against such an adversary who can manipulate attributes of HTML elements, we randomize HTML element `id` and `class` attributes of all HTML elements in our constructed graph. As shown in Figure 11, we find that AdGRAPH’s precision and recall is not impacted at all by randomization of HTML element `id` and `class` attributes. AdGRAPH is robust to HTML element obfuscation attacks because it mainly relies on structural properties of the constructed graph and not on HTML element attributes.

4.4.2 HTTP URL Obfuscation Attacks. Second, we implement an adversarial publisher who can manipulate HTTP URLs. HTTP URL blocking rules in filter lists typically use domain and query string information. Thus, an adversarial publisher can simply randomize domain and query string information to bypass these HTTP URL blocking rules. While modifications to query string can generally be easily done without impacting the site’s appearance and functionality, modifications to domain names of URLs are constrained. For example, a third-party advertiser cannot arbitrarily modify its domain to that of the first-party publisher or other publishers. As another example, due to the non-trivial overheads of domain registration and maintenance, a third-party advertiser can only dynamically select its domain from a pool of a few domains. To evaluate the robustness of AdGRAPH against such an adversary, as discussed next, we manipulate information in HTTP URLs in three ways: (1) modify query string, (2) modify domain name, and (3) modify both query string and domain name.

Query string randomization. We modify query string of a URL by randomly changing its parameter names, parameter values, and the number of parameters. For each URL in our constructed graph, we conduct a randomly selected combination of these modifications. Recall from Section 3.4 that we do use query string information to extract features. For example, we look for specific ad-related keywords. An adversary who randomizes query strings would be able to successfully manipulate ad-related keywords. Figure 11 shows that an adversary employing query string randomization has a small impact on precision and recall. Specifically, precision and recall of AdGRAPH decrease by 0.01 and 0.02, respectively.

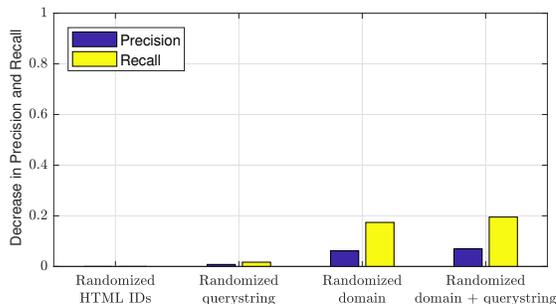


Figure 11: Change in precision and recall of AdGRAPH against different obfuscation attacks.

Domain name randomization. We modify domain name of a URL by adding a random sub-domain (if not already there), randomly changing an existing sub-domain, and randomly changing the base domain. For each URL in our constructed graph, we conduct a randomly selected combination of these modifications. Note that we only add or modify sub-domains for first-party URLs because changing its base domain would make them third-party URLs. Similarly, we do not change the base domain of a third-party URL to that of the first-party. Recall from Section 3.4 that we do use domain information to extract features. For example, we use sub-domain information of HTTP URLs. An adversary who randomizes domain information would be able to successfully manipulate these domain features. Figure 11 shows that an adversary employing domain name randomization has a modest impact on precision and recall. Specifically, precision and recall of AdGRAPH decrease by 0.06 and 0.17, respectively.

Randomization of both query string and domain name. We also jointly modify both query string and domain name for each URL. Figure 11 shows that joint randomization of query string and domain name has the most impact on precision and recall of AdGRAPH. Specifically, precision and recall of AdGRAPH decrease by 0.07 and 0.19, respectively.

Overall, we conclude that AdGRAPH is fairly robust against adversaries who have the capability of manipulating HTML and HTTP information on web pages. AdGRAPH achieves robustness because it does not rely on any singular source of information (e.g., patterns in query string or blacklist of domains). The graph representation of interactions across HTTP, HTML, and JavaScript layers of the web stack means that an adversary would have to manipulate information in all three layers together to bypass AdGRAPH.

5 RELATED WORK

In this section, we review prior work that use machine learning to replace or complement crowdsourced filter lists for blocking ads, trackers, and anti-adblockers.

HTTP-based approaches. Researchers have previously attempted to identify patterns in HTTP requests using supervised machine learning approaches to aid with filter list maintenance. Bhagavatula et al. [30] used keywords and query string information in HTTP requests to train supervised machine learning models for

classifying ads. Gugelmann et al. [38] used HTTP request header attributes such as number of HTTP requests and size of payload to identify tracking domains. Yu et al. [57] also analyzed HTTP requests to detect privacy leaks by trackers in URL query string parameters. While these approaches are somewhat automated, they are not robust to evasion, due to their reliance on simple HTTP-based features that can be easily evaded by adversaries (publishers or advertisers) who can manipulate domain or query string information.

HTML-based approaches. Researchers have also attempted to identify HTML DOM patterns using computer vision and machine learning techniques to detect ads and anti-adblockers. For example, Storey et al. [49] proposed a perceptual adblocking approach to detect ads based on their visual properties. Their key insight is that ads need to be distinguishable from organic content as per government regulations [18] and industry self-regulations [23]. Mughees et al. [45] analyzed patterns in DOM changes to train a machine learning model for detecting anti-adblockers. These approaches are also not robust to evasion due to their reliance on simple HTML-based features. Adversarial publishers and advertisers can easily manipulate HTML element attributes to defeat these approaches.

JavaScript-based approaches. Since publishers and advertisers extensively rely on JavaScript to implement advertising and tracking functionalities, researchers have also tried to use JavaScript code analysis for detecting ads and trackers. To this end, one thread of research leverages static analysis of JavaScript code using machine learning techniques. Ikram et al. [39] conducted n-gram analysis of JavaScript dependency graphs using one-class support vector machine for detecting trackers. Iqbal et al. [40] constructed abstract syntax trees of JavaScript code which were then mapped to features for training machine learning models to detect anti-ad blockers. While these static analysis approaches achieve good accuracy, they are not robust against even simple JavaScript code obfuscation techniques [50, 55, 56]. Researchers have resorted to dynamic analysis techniques to overcome some of these challenges. Wu et al. [53] extracted JavaScript API invocations through dynamic analysis and trained a machine learning model to detect trackers. Storey et al. [49] proposed to intercept and modify API calls by JavaScript to bypass anti-adblockers. Zhu et al. [60] conducted differential JavaScript execution analysis to identify branch statements such as if/else that are triggered by anti-adblocking scripts. While dynamic analysis is more resilient to obfuscation than static analysis, it is possible for websites to conceal and obfuscate JavaScript APIs to reduce the effectiveness of dynamic analysis.

Multi-layer approaches. To improve the accuracy and robustness of ad blocking, researchers have looked at integrating multiple information types (e.g., HTML, HTTP, JavaScript). Bau et al. [29] proposed to leverage both HTML and HTTP information with machine learning to automatically capture relationships among web content to robustly detect evasive trackers. More specifically, the authors constructed the DOM tree, labeled each node with its domain, and then extracted a wide range of graph properties (e.g., depth, degree) for each domain. Kaizer and Gupta [41] utilized both JavaScript and HTTP information to train a machine learning model for detecting trackers. More specifically, the authors used JavaScript navigation and screen properties such as `appName` and `plugins` and

HTTP attributes, such as cookies and URL length to detect tracking URLs. Our proposed approach `ADGRAPH` significantly advances this line of research. To the best of our knowledge, `ADGRAPH` represents the first attempt to comprehensively capture interactions between HTML, HTTP, and JavaScript on a web page to detect ads and trackers. As our evaluations have shown, leveraging all of the available information at multiple layers helps `ADGRAPH` in accurately and robustly detecting ads and trackers.

6 CONCLUSION

We presented a graph-based machine learning approach, called `ADGRAPH`, to automatically and effectively block ads and trackers on the web. The key insight behind `ADGRAPH` is to leverage information obtained from multiple layers of the web stack: HTML, HTTP, and JavaScript. With these three ingredients brought together, we showed that we can train supervised machine learning models to automatically block ads and trackers. We found that `ADGRAPH` replicates the behavior of popular crowdsourced filter lists with an 97.7% accuracy. In addition, `ADGRAPH` is able to detect a significant number of ads and tracker which are missed by popular crowdsourced filter lists.

In summary, `ADGRAPH` represents a significant advancement over unreliable crowdsourced filter lists that are used by state-of-the-art adblocking browsers and extensions. More importantly, as adblockers pose a growing threat to the ad-driven “free” web, we expect more and more financially motivated publishers and advertisers to employ adversarial obfuscation techniques to evade adblockers. Unfortunately, crowdsourced filter lists used by state-of-the-art adblockers can be easily evaded using simple obfuscation techniques. Therefore, `ADGRAPH`’s resistance to adversarial obfuscation attempts by publishers and advertisers represents an important technical advancement in the rapidly escalating adblocking arms race.

REFERENCES

- [1] Adblock Plus for Chrome support. <https://adblockplus.org/forum/viewforum.php?f=10>.
- [2] Anti-Adblock Killer. <https://github.com/reek/anti-adblock-killer>.
- [3] Anti-Adblock Killer List Forum. <https://github.com/reek/anti-adblock-killer/issues>.
- [4] Blockzilla. <https://zpacman.github.io/Blockzilla/>.
- [5] Digitrust, - neutral, non-profit, standardized id for digital media. <http://www.digitru.st/>.
- [6] EasyList. <https://easylist.to/>.
- [7] EasyList Forum. <https://forums.lanik.us>.
- [8] EasyList variants. <https://easylist.to/pages/other-supplementary-filter-lists-and-easylist-variants.html>.
- [9] easylist/easylist commits. <https://github.com/easylist/easylist/commits/master>.
- [10] EasyPrivacy. <https://easylist.to/easylist/easylist.txt>.
- [11] Fanboy Annoyances List. <https://www.fanboy.co.nz/>.
- [12] Fanboy Social List. <https://www.fanboy.co.nz/>.
- [13] Font awesome. <https://fontawesome.com/>.
- [14] Instart Logic AppShield Ad Integrity. <https://www.instartlogic.com/products/advertising-marketing-recovery>.
- [15] PageFair, 2017 Global Adblock Report. <https://pagefair.com/downloads/2017/01/PageFair-2017-Adblock-Report.pdf>.
- [16] Peter Lowe’s list. <http://pgl.yoyo.org/adservers/>.
- [17] Squid blacklist. <http://www.squidblacklist.org/>.
- [18] Truth In Advertising, Federal Trade Commission. <https://www.ftc.gov/news-events/media-resources/truth-advertising/>.
- [19] Twitter emoji (twemoji). <https://github.com/twitter/twemoji>.
- [20] uBlockOrigin/uAssets commits. <https://github.com/uBlockOrigin/uAssets/commits/master>.

- [21] Warning removal list. <https://easylist-downloads.adblockplus.org/antiadblockfilters.txt>.
- [22] yimg.com conversiontrackerservice. <https://github.com/yahoojp-marketing/sponsored-search-api-documents/blob/master/docs/en/api,eference/services/ConversionTrackerService.md>.
- [23] YourAdChoices. <http://youradchoices.com/>.
- [24] PageFair 2015 Adblock Report. <https://pagefair.com/blog/2015/ad-blocking-report/>, 2015.
- [25] An experimental methodology to measure consumers' perceptions of online ad experiences. Ad Experience Research Group, Coalition for Better Ads, 2016.
- [26] PageFair 2016 Mobile Adblocking Report. <https://pagefair.com/blog/2016/mobile-adblocking-report/>, 2016.
- [27] Determining a Better Ads Standard Based on User Experience Data. Coalition for Better Ads, 2017. <https://www.betterads.org/wp-content/uploads/2017/03/Determining-a-Better-Ads-Standard-based-on-User-Experience-Data.pdf>.
- [28] G. Anthes. Data Brokers Are Watching You. *Communications of the ACM*, 2015.
- [29] J. Bau, J. Mayer, H. Paskov, and J. C. Mitchell. A Promising Direction for Web Tracking Countermeasures. In *W2SP*, 2013.
- [30] S. Bhagavatula, C. Dunn, C. Kanich, M. Gupta, and B. Ziebart. Leveraging Machine Learning to Improve Unwanted Resource Filtering. In *ACM Workshop on Artificial Intelligence and Security*, 2014.
- [31] L. Bilge and T. Dumitras. Before we knew it: an empirical study of zero-day attacks in the real world. In *ACM Conference on Computer and Communications Security (CCS)*, 2012.
- [32] J. Bloomberg. Ad Blocking Battle Drives Disruptive Innovation. *Forbes*, 2017. <https://www.forbes.com/sites/jasonbloomberg/2017/02/18/ad-blocking-battle-drives-disruptive-innovation/#156729bd601e>.
- [33] A. Bosworth. A New Way to Control the Ads You See on Facebook, and an Update on Ad Blocking. <https://newsroom.fb.com/news/2016/08/a-new-way-to-control-the-ads-you-see-on-facebook-and-an-update-on-ad-blocking/>, 2016.
- [34] L. Breiman. Random Forests. In *Machine learning*, 2001.
- [35] C. Cimpanu. Ad Network Uses DGA Algorithm to Bypass Ad Blockers and Deploy In-Browser Miners. <https://www.bleepingcomputer.com/news/security/ad-network-uses-dga-algorithm-to-bypass-ad-blockers-and-deploy-in-browser-miners/>, 2018.
- [36] S. Englehardt and A. Narayanan. Online Tracking: A 1-million-site Measurement and Analysis. In *ACM Conference on Computer and Communications Security (CCS)*, 2016.
- [37] S. Eskandari, A. Leoutsarakos, T. Mursch, and J. Clark. A first look at browser-based cryptojacking. In *IEEE Security & Privacy on the Blockchain*, 2018.
- [38] D. Gugelmann, M. Happe, B. Ager, and V. Lenders. An Automated Approach for Complementing Ad Blockers' Blacklists. In *Privacy Enhancing Technologies Symposium (PETS)*, 2015.
- [39] M. Ikram, H. J. Asghar, M. A. Kaafar, A. Mahanti, and B. Krishnamurthy. Towards Seamless Tracking-Free Web: Improved Detection of Trackers via One-class Learning. In *Privacy Enhancing Technologies Symposium (PETS)*, 2017.
- [40] U. Iqbal, Z. Shafiq, and Z. Qian. The Ad Wars: Retrospective Measurement and Analysis of Anti-Adblock Filter Lists. In *IMC*, 2017.
- [41] A. J. Kaizer and M. Gupta. Towards Automatic identification of JavaScript-oriented Machine-Based Tracking. In *IWSPA*, 2016.
- [42] O. Katz and B. Livshits. Toward an evidence-based design for reactive security policies and mechanisms. Technical Report CS-2016-04-2016, Technion, Nov. 2016.
- [43] B. Li, P. Vadrevu, K. H. Lee, and R. Perdisci. JSgraph: Enabling Reconstruction of Web Attacks via Efficient Tracking of Live In-Browser JavaScript Executions. In *NDSS*, 2018.
- [44] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang. Knowing your enemy: understanding and detecting malicious web advertising. In *ACM Conference on Computer and Communications Security (CCS)*, 2012.
- [45] M. H. Mughees, Z. Qian, and Z. Shafiq. Detecting Anti Ad-blockers in the Wild. In *Privacy Enhancing Technologies Symposium (PETS)*, 2017.
- [46] R. Nithyanand, S. Khattak, M. Javed, N. Vallina-Rodriguez, M. Falahrestegar, J. E. Powles, E. D. Cristofaro, H. Haddadi, and S. J. Murdoch. Adblocking and Counter-Blocking: A Slice of the Arms Race. In *USENIX Workshop on Free and Open Communications on the Internet*, 2016.
- [47] G. Sloane. Ad Blocker's Successful Assault on Facebook Enters Its Second Month. <http://adage.com/article/digital/blockrace-adblock/311103/>, 2017.
- [48] A. K. Sood and R. J. Enbody. Malvertising – exploiting web advertising. *Computer Fraud & Security*, 2011.
- [49] G. Storey, D. Reisman, J. Mayer, and A. Narayanan. The Future of Ad Blocking: An Analytical Framework and New Techniques. In *arXiv:1705.08568*, 2017.
- [50] Z. Tzermias, G. Sykiotakis, M. Polychronakis, and E. P. Markatos. Combining static and dynamic analysis for the detection of malicious documents. In *Fourth European Workshop on System Security (EUROSEC)*, 2011.
- [51] W. Wang, Y. Zheng, X. Xing, Y. Kwon, X. Zhang, and P. Eugster. WebRanz: Web Page Randomization For Better Advertisement Delivery and Web-Bot Prevention. In *ACM SIGSOFT International Symposium on Foundations of Software Engineering*, 2016.
- [52] B. Williams. Ping pong with Facebook. <https://adblockplus.org/blog/ping-pong-with-facebook>, 2018.
- [53] Q. Wu, Q. Liu, Y. Zhang, P. Liu, and G. Wen. A Machine Learning Approach for Detecting Third-Party Trackers on the Web. In *ESORICS*, 2016.
- [54] X. Xing, W. Meng, B. Lee, U. Weinsberg, A. Sheth, R. Perdisci, and W. Lee. Understanding Malvertising Through Ad-Injecting Browser Extensions. In *International Conference on World Wide Web (WWW)*, 2015.
- [55] W. Xu, F. Zhang, and S. Zhu. The power of obfuscation techniques in malicious JavaScript code: A measurement study. In *7th International Conference on Malicious and Unwanted Software (MALWARE)*, 2012.
- [56] W. Xu, F. Zhang, and S. Zhu. JStill: Mostly Static Detection of Obfuscated Malicious JavaScript Code. In *ACM Conference on Data and Application Security and Privacy (CODASPY)*, 2013.
- [57] Z. Yu, S. Macbeth, K. Modi, and J. M. Pujol. Tracking the Trackers. In *World Wide Web (WWW) Conference*, 2016.
- [58] Z. Zaifeng. Who is Stealing My Power III: An Adnetwork Company Case Study, 2018. <http://blog.netlab.360.com/who-is-stealing-my-power-iii-an-adnetwork-company-case-study-en/>.
- [59] C. Zhang and Y. Ma, editors. *Ensemble Machine Learning*. Springer-Verlag New York, 2012.
- [60] S. Zhu, X. Hu, Z. Qian, Z. Shafiq, and H. Yin. Measuring and Disrupting Anti-Adblockers Using Differential Execution Analysis. In *NDSS*, 2018.