sciendo

Proceedings on Privacy Enhancing Technologies ; 2018 (4):85–103

Muhammad Ahmad Bashir and Christo Wilson

Diffusion of User Tracking Data in the Online Advertising Ecosystem

Abstract: Advertising and Analytics (A&A) companies have started collaborating more closely with one another due to the shift in the online advertising industry towards Real Time Bidding (RTB). One natural way to understand how user tracking data moves through this interconnected advertising ecosystem is by modeling it as a graph. In this paper, we introduce a novel graph representation, called an *Inclusion* graph, to model the impact of RTB on the diffusion of user tracking data in the advertising ecosystem. Through simulations on the *Inclusion* graph, we provide upper and lower estimates on the tracking information observed by A&A companies. We find that 52 A&A companies observe at least 91% of an average user's browsing history under reasonable assumptions about information sharing within RTB auctions. We also evaluate the effectiveness of blocking strategies (e.g., AdBlock Plus), and find that major A&A companies still observe 40–90% of user impressions, depending on the blocking strategy.

Keywords: Online Tracking, RTB, Cookie Matching

DOI 10.1515/popets-2018-0033
Received 2018-02-28; revised 2018-06-15; accepted 2018-06-16.

1 Introduction

In the last decade, the online display advertising industry has massively grown in size and scope. According to the Interactive Advertising Bureau (IAB), revenue from the online display ad industry in the U.S. totaled $88B in 2017, a growth of 21.4% from 2016 [63]. This increased spending is fueled by advances that enable advertisers to target users with increasing levels of precision, even across different devices and platforms.

Another recent change in the online display advertising ecosystem is the shift from *ad networks* to *ad exchanges*, where advertisers bid on *impressions* being sold in Real Time Bidding (RTB) auctions. The rise of RTB has forced Advertising and Analytics (*A&A*) companies to collaborate more closely with one another, in order to exchange data about users and facilitate bidding on impressions [10, 58]. The move towards RTB has also caused A&A companies to specialize into particular roles. For example, Supply-Side Platforms (SSPs) work with *publishers* (e.g., CNN) to help manage their relationship with ad exchanges, while Demand-Side Platforms (DSPs) try to optimize ad placement and bidding on behalf of advertisers. In short, due to RTB, the online advertising ecosystem has become enormously complex.

A natural way to model this complex ecosystem is in the form of a graph. Graph models that accurately capture the relationships between publishers and A&A companies are extremely important for practical applications, such as estimating revenue of A&A companies [26], predicting whether a given domain is a tracker [34], or evaluating the effectiveness of domain-blocking strategies on preserving users' privacy.

However, to date, technical limitations have prevented researchers from developing accurate graph models of the online advertising ecosystem. For example, Gomer et al. [29] propose a *Referer* graph, where nodes represent publishers or A&A domains, and two nodes a_i and a_j are connected if an HTTP message to a_j is observed with a_i as the HTTP `Referer`. Unfortunately, as we will show, graphs built using `Referer` information may contain erroneous edges in cases where a third-party script is embedded directly into a first-party context (i.e., is not sandboxed in an `iframe`).

In this paper, to model the diffusion of user tracking data within RTB auctions, we propose a novel and accurate representation of the advertising graph called an *Inclusion* graph. The *Inclusion* graph corrects the technical problem of the *Referer* graph by using the actual inclusion relationships between domains to represent edges, rather than imprecise `Referer` relationships. We are able to construct *Inclusion* graphs, thanks to advances in browser instrumentation that allow researchers to conduct web crawls that record the exact provenance of all HTTP(S) requests [6, 10, 41].

We use crawled data consisting of around 2M impressions from popular e-commerce websites collected

Muhammad Ahmad Bashir: Northeastern University, E-mail: ahmad@ccs.neu.edu
Christo Wilson: Northeastern University, E-mail: cbw@ccs.neu.edu

by a specially instrumented version of Chrome [10] to construct the *Inclusion* graph. In § 4, we examine the fundamental graph properties of the *Inclusion* graph and compare it to a *Referer* graph, created using the same dataset to understand their salient differences. In § 5, we demonstrate a concrete use case for the *Inclusion* graph by using simulations to model the flow of tracking data to A&A companies. Furthermore, we compare the efficacy of different real-world and graph theoretic "blocking" strategies (e.g., AdBlock Plus [2], Ghostery [25], and Disconnect [18]) at reducing the flow of tracking information to A&A companies.

Overall, we make the following key contributions:

- We introduce the *Inclusion* graph as a model for capturing the complexity of the online advertising ecosystem. We use the *Inclusion* graph as a substrate for modeling the flow of impressions to A&A companies by taking into account the browsing behavior of users and the dynamics of RTB auctions.
- We find that the *Inclusion* graph has substantive differences in graph structure compared to the *Referer* graph because 48.4% of resource inclusions in our crawled data have an inaccurate `Referer`.
- Through simulations, we find that 52 A&A companies are each able to observe 91% of an average user's impressions as they browse, under modest assumptions about data sharing in RTB auctions. 636 A&A companies are able to observe at least 50% of an average user's impressions. Even under the strictest simulation assumptions, the top 10 A&A companies observe 89-99% of all user impressions.
- We simulate the effect of five blocking strategies, and find that AdBlock Plus (the world's most popular ad blocking browser extension [45, 62], is ineffective at protecting users' privacy because major ad exchanges are whitelisted under the Acceptable Ads program [73]. In contrast, Disconnect blocks the most information flows to A&A companies, followed by removal of top 10% A&A nodes. However, even with strong blocking, major A&A companies still observe 40–80% of user impressions.

The raw data we use in this study is publicly available.[1] We have also publicly released the source code and data from this study.[2]

1 http://personalization.ccs.neu.edu/Projects/Retargeting/

2 http://personalization.ccs.neu.edu/Projects/AdGraphs/

2 Background and Related Work

In this section, we review technical details of and current computer science research on the online display advertising ecosystem. We start by discussing related work on user privacy and tracking. Next, we present examples of the current display ad serving process and define the roles of different actors in the ecosystem, followed by a brief overview of efforts to empirically measure these processes. Lastly, we examine prior work that modeled the ad ecosystem as a graph.

2.1 Tracking and Blocking

To show relevant ads to users, advertisers rely heavily on collecting information about users as they browse the web. This data collection is achieved by embedding trackers into webpages that gather browsing information about each user.

The area of tracking has been well studied. Krishnamurthy et al. and others have documented the pervasiveness of trackers and the associated user privacy implications over time [15, 20, 26, 33, 37–39]. Furthermore, tracking techniques have evolved over time. Persistent cookies [35], local state in browser plugins [7, 68, 69], and various browser fingerprinting methods [1, 21, 36, 51, 55, 57, 65] are some of the techniques that have been deployed to track users. Englehardt et al. [20] found evidence of tracking via the `Audio` and `Battery Status` JavaScript APIs. In addition to tracking users themselves, advertisers try to maximize their knowledge of each user's interest profile by sharing information with each other via cookie matching [1, 10, 23, 58]. Falahrastegar et al. examine how tracking differs across geographic regions [22].

Users have become increasingly concerned with the amount and types of tracking information collected about them [47, 70]. Several surveys have investigated users' concerns about targeted ads, their preferences towards tracking, and usage of privacy tools [8, 42, 48, 66, 71]. Concerns about the privacy implications of tracking (as well as the insecurity of online ad networks [75]) has led to increased adoption of tools that block trackers and ads. Two studies have examined the usage of ad blockers in-the-wild [45, 62], while Walls et al. looked at efforts to whitelist "acceptable advertisers" [73].

Merzdovnik et al. critically examined the effectiveness of tracker blocking tools [49]; in contrast, Nithyanand et al. studied advertisers' efforts to counter

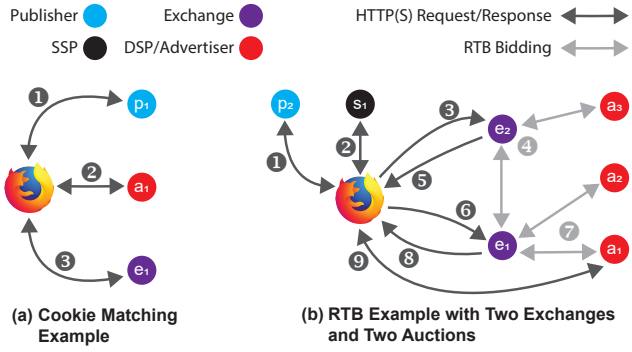


Fig. 1. Examples of **(a)** cookie matching and **(b)** showing an ad to a user via RTB auctions. **(a)** The user visits publisher p_1 ❶ which includes JavaScript from advertiser a_1 ❷. a_1's JavaScript then cookie matches with exchange e_1 by programmatically generating a request that contains both of their cookies ❸. **(b)** The user visits publisher p_2, which then includes resources from SSP s_1 and exchange e_2 ❶–❸. e_2 solicits bids ❹ and sells the impression to e_1 ❺ ❻, which then holds another auction ❼, ultimately selling the impression to a_1 ❽ ❾.

ad blockers [56]. Mughees et al. examined the prevalence of anti-ad blockers in the wild [53]. In this work, we expand on the existing blocking literature by taking the effects of ad auctions and cookie matching into account.

The research community has proposed a variety of mechanisms to stop online tracking that go beyond blacklists of domains and URLs. Li et al. [43] and Ikram et al. [32] used machine learning to identify trackers, while Papaodyssefs et al. [60] examined the use of private cookies to avoid being tracked. Nikiforakis et al. propose the complementary idea of adding entropy to the browser to evade fingerprinting [54]. However, despite these efforts, third-party trackers are still pervasive and pose real privacy issues to users [49].

2.2 The Online Advertising Ecosystem

Numerous studies have chronicled the online advertising ecosystem, which is composed of companies that: track users, serve ads, act as platforms between *publishers* (websites that rely on advertising revenue to pay for content creation) and advertisers, or all of the above. Mayer et al. present an accessible introduction to this topic in [46]. **In this work, we collectively refer to companies engaged in analytics and advertising as *A&A* companies.**

Recently, the online ad ecosystem has begun to shift from ad networks to *ad exchanges*, which implement Real Time Bidding (RTB) auctions to sell *impressions* to advertisers. In the advertising industry, the term "impression" is used when advertising or tracking content is rendered in a user's browser after they visit a webpage [17]. To participate in RTB auctions, A&A companies must implement *cookie matching*, which is a process by which different A&A companies exchange their unique tracking identifiers for specific users. Several studies have examined the emergence of cookie matching [1, 10, 23, 58]. Ghosh et al. theoretically model the incentives for A&A companies to collaborate with their competitors in RTB auction systems [24].

Figure 1(a) illustrates the typical process used by A&A companies to match cookies. When a user visits a website ❶, JavaScript code from a third-party advertiser a_1 is automatically downloaded and executed in the user's browser ❷. This code may set a cookie in the user's browser, but this cookie will be unique to a_1, i.e., it will not contain the same unique identifiers as the cookies set by any other A&A companies. Furthermore, the Same Origin Policy (SOP) prevents a_1's code from reading the cookies set by any other domain. To facilitate bidding in future RTB auctions, a_1 must match its cookie to the cookie set by an ad exchange like e_1. As shown in the figure, a_1's JavaScript accomplishes this by programmatically causing the browser to send a request to e_1 ❸. The JavaScript includes a_1's cookie in the request, and the browser automatically adds a copy of e_1's cookie, thus allowing e_1 to create a match between its cookie and a_1's.

Figure 1(b) shows an example of how an ad may be shown on publisher p_2 using RTB auctions. When a user visits p_2 ❶, JavaScript code is automatically downloaded and executed either from a *Supply Side Platform (SSP)* ❷ or an ad exchange. SSPs are A&A companies that specialize in maximizing publisher revenue by forwarding impressions to the most lucrative ad exchange. Eventually the impression arrives at the auction held by ad exchange e_2 ❸, and e_2 solicits bids from advertisers and *Demand Side Platforms (DSPs)* ❹. DSPs are A&A companies that specialize in executing ad campaigns on behalf of advertisers. Note that **all participants in the auction observe the impression**; however, because only e_2's cookie is available at this point, auction participants that have not matched cookies with e_2 will not be able to identify the user.

The process of filling an impression may continue even after an RTB auction is won, because the winner may be yet another ad exchange or ad network. As shown in Figure 1(b), the impression is purchased from e_2 by e_1 ❺ ❻, who then holds another auction ❼ and ultimately sells to a_1 (the advertiser from the cookie matching example) ❽ ❾. Ad exchanges and ad networks

routinely match cookies with each other to facilitate the flow of impression inventory between markets.

Measurement Studies. Barford et al. broadly characterized the web *adscape* and identified systematically important ad networks [9]. Rodriguez et al. measured the ad ecosystem that serves mobile devices [72], while Zarras et al. specifically examined ad networks that serve malicious ads [75]. Gill et al. modeled the revenue earned by different A&A companies [26], while other studies have used empirical measurements to determine the value of individual users to online advertisers [58, 59]. Many studies have used a variety of methods to study the targeted ads that are displayed to users under a variety of circumstances [9–11, 16, 30, 44].

2.3 Ad Ecosystem Graphs

A natural structure for modeling the online ad ecosystem is a graph, where nodes represent publishers and A&A companies, and edges capture relationships between these entities. Gomer et al. [29] built and analyzed graphs of the ad ecosystem by making use of the *Referer* field from HTTP requests. In this representation, a relationship $d_i \rightarrow d_j$ exists if there is an HTTP request to domain d_j with a *Referer* header from domain d_i.

While Gomer et al. provided interesting insights into the structure of the ad ecosystem, their referral-based graph representation has a significant limitation. As we describe in § 3.3, relying on the HTTP *Referer* **does not** always capture the correct relationships between A&A parties, thus leading to incorrect graphs of the ad ecosystem. We re-create this graph representation using our dataset (see § 3) and compare its properties to a more accurate representation in § 4.

Kalavri et al. [34] created a bipartite graph of publishers and associated A&A domains, then transformed it to create an undirected graph consisting solely of A&A domains. In their representation, two A&A domains are connected if they were included by the same publisher. This construction leads to a highly dense graph with many complete cliques. Kalavri et al. leveraged the tight community structure of A&A domains to predict whether new, unknown URLs were A&A or not. However, this co-occurrence representation has a conceptual shortcoming: it may include edges between A&A domains that do not directly communicate or have any business relationship. Due to this shortcoming, we do **not** explore this graph representation in this work.

3 Methodology

Our goal is to capture the most accurate representation of the online advertising ecosystem, which will allow us to model the effect of RTB on diffusion of user tracking data. In this section, we introduce the dataset used in this study and describe how we use it to build a graph representation of the ad ecosystem.

3.1 Dataset

In this work, we use the dataset provided by Bashir et al. [10]. The goal of [10] was to causally infer the information sharing relationships between A&A companies by (1) crawling products from popular e-commerce websites and then (2) observing corresponding *retargeted* ads on publishers. Bashir et al. conducted web crawls that covered 738 major e-commerce websites (e.g., Amazon) and 150 popular publishers (e.g., CNN).[3] The authors chose top e-commerce sites from Alexa's hierarchical list of online shops [4], and manually chose publishers from the Alexa Top-1K. They crawled 10 manually selected products per e-commerce site to signal strong *intent* to trackers and advertisers, followed by 15 randomly chosen pages per publisher to elicit display ads. In total, Bashir et al. repeated the entire crawl nine times, resulting in data for around 2M impressions.

3.2 Inclusion Trees

Bashir et al. [10] used a specially instrumented version of Chromium for their web crawls. Their crawler recorded the *inclusion tree* for each webpage, which is a data structure that captures the semantic relationships between elements in a webpage (as opposed to the DOM, which captures syntactic relationships) [6, 41]. The crawler also recorded all HTTP request and response headers associated with each visited URL.

To illustrate the importance of inclusion trees, consider the example webpage shown in Figure 2(a). The DOM shows that the page from publisher p ultimately includes resources from four third-party domains (a_1 through a_4). It is clear from the DOM that the request to a_3 is responsible for causing the request to a_4, since the `script` inclusion is within the `iframe`. However, it

3 For simplicity, we refer to these e-commerce websites as publishers, to distinguish them from A&A domains.

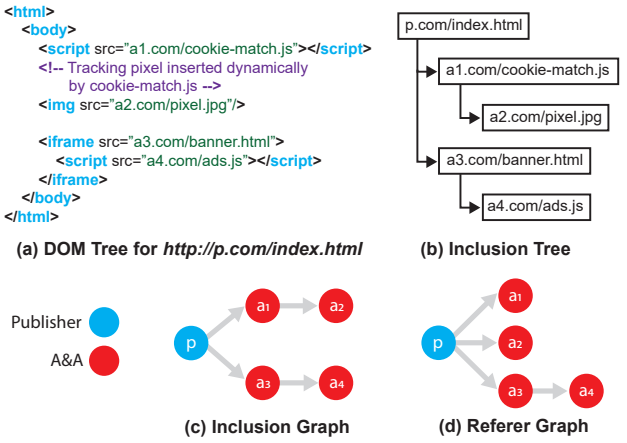


Fig. 2. An example HTML document and the corresponding inclusion tree, *Inclusion* graph, and *Referer* graph. In the DOM representation, the a_1 `script` and a_2 `img` appear at the same level of the tree; in the inclusion tree, the a_2 `img` is a child of the a_1 `script` because the latter element created the former. The *Inclusion* graph has a 1:1 correspondence with the inclusion tree. The *Referer* graph fails to capture the relationship between the a_1 `script` and a_2 `img` because they are both embedded in the first-party context, while it correctly attributes the a_4 `script` to the a_3 `iframe` because of the context switch.

is not clear which domain generated the requests to a_2 and a_3: the `img` and `iframe` could have been embedded in the original HTML from p, or these elements could have been created dynamically by the `script` from a_1. In this case, the inclusion tree shown in Figure 2(b) reveals that the `image` from a_2 was dynamically created by the `script` from a_1, while the `iframe` from a_3 was embedded directly in the HTML from p.

The instrumented Chromium binary used by Bashir et al. was able to correctly determine the provenance of webpage elements, regardless of how they were created (e.g., directly in HTML, via inline or remotely included `script` tags, dynamically via `eval()`, etc.), or where they were located (in the main context or within `iframes`). This was accomplished by tagging all scripts with provenance information (i.e., first-party for inline scripts), and then dynamically monitoring the execution of each script. New scripts created during the execution of a given script (e.g., via `document.write()`) were linked to their parent.[4] More details about how Chromium was instrumented and inclusion trees were extracted are available in [6].

Cookie Matching. The Bashir et al. dataset also includes labels on edges of the inclusion trees indicating cases where cookie matching is occurring. These labels are derived from heuristics (e.g., string matching to identify the passing of cookie values in HTTP parameters) and causal inferences based on the presence of retargeted ads. We use this data in § 5 to constrain some of our simulations.

3.3 Graph Construction

A natural way to model the online ad ecosystem is using a graph. In this model, nodes represent A&A companies, publishers, or other online services. Edges capture relationships between these actors, such as resource inclusion or information flow (e.g., cookie matching).

Canonicalizing Domains. We use the data described in § 3.1 to construct a graph for the online advertising ecosystem. We use effective 2^{nd}-level domain names to represent nodes. For example, `x.doubleclick.net` and `y.doubleclick.net` are represented by a single node labeled `doubleclick`. Throughout this paper, when we say "domain", we are referring to an effective 2^{nd}-level domain name.[5]

Simplifying domains to the effective 2^{nd}-level is a natural encoding for advertising data. Consider two inclusion trees generated by visiting two publishers: publisher p_1 forwards the impression to `x.doubleclick.net` and then to advertiser a_1. Publisher p_2 forwards to `y.doubleclick.net` and advertiser a_2. This does not imply that `x.doubleclick` and `y.doubleclick` only sell impressions to a_1 and a_2, respectively. In reality, DoubleClick is a single auction, regardless of the subdomain, and a_1 and a_2 have the opportunity to bid on all impressions. Individual inclusion trees are snapshots of how one particular impression was served; only in aggregate can all participants in the auctions be enumerated. Further, 3^{rd}-level domains may read 2^{nd}-level cookies without violating the Same Origin Policy [52]: `x.doubleclick.com` and `y.doubleclick.com` may both access cookies set by `.doubleclick`, and do in practice.

The sole exception to our domain canonicalization process is Amazon's Cloudfront Content Delivery Network (CDN). We routinely observed Cloudfront hosting ad-related scripts and images in our data. We manually examined the 50 fully-qualified Cloudfront domains

4 Note that JavaScript within a given page context executes serially, so there is no ambiguity created by concurrency. Although Web Workers may execute concurrently, they cannot include third party scripts or modify the DOM.

5 None of the publishers and A&A domains in our dataset have two-part TLDs, like `.co.uk`, which simplifies our analysis.

(e.g., d31550gg7drwar.cloudfront.net) that were pre- or proceeded by A&A domains in our data, and mapped each one to the corresponding A&A company (e.g., adroll in this case).

***Inclusion* graph.** We propose a novel representation called an *Inclusion* graph that is the union of all inclusion trees in our dataset. Our representation is a directed graph of publishers and A&A domains. An edge $d_i \rightarrow d_j$ exists if we have ever observed domain d_i including a resource from d_j. Edges may exist from publishers to A&A domains, or between A&A domains. Figure 2(c) shows an example *Inclusion* graph.

***Referer* graph.** Gomer et al. [29] also proposed a directed graph representation consisting of publishers and A&A domains for the online advertising ecosystem. In this representation, each publisher and A&A domain is a node, and edge $d_i \rightarrow d_j$ exists if we have ever observed an HTTP request to d_j with Referer d_i. Figure 2(d) shows an example *Referer* graph corresponding to the given webpage. The Bashir et al. [10] dataset includes all HTTP request and response headers from the crawl, and we use these to construct the *Referer* graph.

Although the *Referer* and *Inclusion* graphs seem similar, they are fundamentally different for technical reasons. Consider the examples shown in Figure 2: the script from a_1 is included directly into p's context, thus p is the Referer in the request to a_2. This results in a *Referer* graph with two edges that does **not** correctly encode the relationships between the three parties: $p \rightarrow a_1$ and $p \rightarrow a_2$. In other words, HTTP Referer headers are an indirect method for measuring the semantic relationships between page elements, and the headers may be incorrect depending on the syntactic structure of a page. Our *Inclusion* graph representation fixes the ambiguity in the *Referer* graph by explicitly relying on the inclusion relationships between elements in webpages. We analyze the salient differences between the *Referer* and *Inclusion* graph in § 4.

Weights. Additionally, we also create a weighted version of these graphs. In the *Inclusion* graph, the weight of $d_i \rightarrow d_j$ encodes the number of times a resource from d_i sent an HTTP request to d_j. In the *Referer* graph, the weight of $d_i \rightarrow d_j$ encodes the number of HTTP requests with Referer d_i and destination d_j.

3.4 Detection of A&A Domains

For us to understand the role of A&A companies in the advertising graph, we must be able to distinguish A&A domains from publishers and non-A&A third parties like CDNs. In the *inclusion trees* from the Bashir et al. dataset [10], each resource is labeled as A&A or non-A&A using the EasyList and EasyPrivacy rule lists. For all the A&A labeled resources, we extract the associated 2^{nd}-level domain. To eliminate false positives, we only consider a 2^{nd}-level domain to be A&A if it was labeled as A&A more than 10% of the time in the dataset.

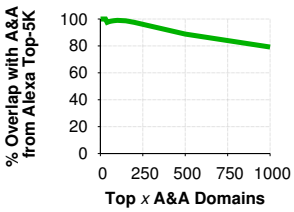


Fig. 3. Overlap between frequent A&A domains and A&A domains from Alexa Top-5K.

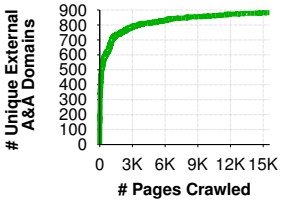


Fig. 4. Unique A&A domains contacted by each A&A domain as we crawl more pages.

3.5 Coverage

There are two potential concerns with the raw data we use in this study: *does the data include a representative set of A&A domains?* and *does the data contain all of the outgoing edges associated with each A&A domain?* To answer the former question, we plot Figure 3, which shows the overlap between the top x A&A domains in our dataset (ranked by inclusion frequency by publishers) with all of the A&A domains included by the Alexa Top-5K websites.[6] We observe that 99% of the 150 most frequent A&A domains appear in both samples, while 89% of the 500 most frequent appear in both. These findings confirm that our dataset includes the vast majority of prominent A&A domains that users are likely to encounter on the web.

To answer the second question, we plot Figure 4, which shows the number of unique external A&A domains contacted by A&A domains in our dataset as the crawl progressed (i.e., starting from the first page crawled, and ending with the last). Recall that the dataset was collected over nine consecutive crawls spanning two weeks of time, each of which visited 9,630 individual pages spread over 888 domains.

We observe that the number of A&A →A&A edges rises quickly initially, going from 0 to 800 in 3,600

6 Our dataset and the Alexa Top-5K data were both collected in December 2015, so they are temporally comparable.

Graph Type	$\|V\|$	$\|E\|$	$\|V_{WCC}\|$	$\|E_{WCC}\|$	Avg. Deg. (In)	Avg. Deg. (Out)	Avg. Path Length	Cluster. Coef.	S^{Δ} [31]	Degree Assort.
Inclusion	1917	26099	1909	26099	13.612	13.612	2.748†	0.472‡	31.254‡	-0.31‡
Referer	1923	41468	1911	41468	21.564	21.564	2.429†	0.235‡	10.040‡	-0.29‡

Table 1. Basic statistics for *Inclusion* and *Referer* graph. We show sizes for the largest WCC in each graph. † denotes that the metric is calculated on the largest SCC. ‡ denotes that the metric is calculated on the undirected transformation of the graph.

crawled pages. Then, the growth slows down, requiring an additional 12,000 page visits to increase from 800 to 900. In other words, almost all A&A edges were discovered by half-way through the very first crawl; eight subsequent iterations of the crawl only uncovered 12.5% more edges. This demonstrates that the crawler reached the point of diminishing returns, indicating that the vast majority of connections between A&A domains that existed at the time are contained in the dataset.

4 Graph Analysis

In this section, we look at the essential graph properties of the *Inclusion* graph. This sets the stage for a higher-level evaluation of the *Inclusion* graph in § 5.

4.1 Basic Analysis

We begin by discussing the basic properties of the *Inclusion* graph, as shown in Table 1. For reference, we also compare the properties with those of *Referer* graph.

Edge Misattribution in the *Referer* graph. The *Inclusion* and *Referer* graph have essentially the same number of nodes, however the *Referer* graph has 159% more edges. We observe that 48.4% of resource inclusions in the raw dataset have an inaccurate `Referer` (i.e., the first-party is the `Referer` even though the resource was requested by third-party JavaScript), which is the cause of the additional edges in the *Referer* graph.

There is a massive shift in the location of edges between the *Inclusion* and *Referer* graph: the number of publisher → A&A edges decreases from 33,716 in the *Referer* graph to 10,274 in the *Inclusion* graph, while the number of A&A → A&A edges increases from 7,408 to 13,546. In the *Referer* graph only 3% of A&A → A&A edges are reciprocal, versus 31% in the *Inclusion* graph. Taken together, these findings highlight the practical consequences of misattributing edges based on `Referer` information, i.e., relationships between A&A companies that should be in the core of the network are incorrectly attached to publishers along the periphery.

Structure and Connectivity. As shown in Table 1, the *Inclusion* graph has large, well-connected components. The largest Weakly Connected Component (WCC) covers all but eight nodes in the *Inclusion* graph, meaning that very few nodes are completely disconnected. This highlights the interconnectedness of the ad ecosystem. The average node degree in the *Inclusion* graph is 13.6, and <7% of nodes have in- or out-degree ≥50. This result is expected: publishers typically only form direct relationships with a small-number of SSPs and exchanges, while DSPs and advertisers only need to connect to the major exchanges. The small number of high-degree nodes are ad exchanges, ad networks, trackers (e.g., Google Analytics), and CDNs.

The *Inclusion* graph exhibits a low average shortest path length of 2.7, and a very high average clustering coefficient of 0.48, implying that it is a "small world" graph. We show the "small-worldness" metric S^{Δ} in Table 1, which is computed for a given undirected graph G and an equivalent random graph G_R[7] as $S^{\Delta} = (C^{\Delta}/C^{\Delta}_R)/(L^{\Delta}/L^{\Delta}_R)$, where C^{Δ} is the average clustering[8] coefficient, and L^{Δ} is the average shortest path length [31]. The *Inclusion* graph has a large $S^{\Delta} \approx 31$, confirming that it is a "small world" graph.

Lastly, Table 1 shows that the *Inclusion* graph is disassortative, i.e., low degree nodes tend to connect to high degree nodes.

Summary. Our measurements demonstrate that the structure of the ad network graph is troubling from a privacy perspective. Short path lengths and high clustering between A&A domains suggest that data tracked from users will spread rapidly to all participants in the ecosystem (we examine this in more detail in § 5). This rapid spread is facilitated by high-degree hubs in the

7 Equivalence in this case means that for G and G_R, $|V| = |V_R|$ and $|E|/|V| = |E_R|/|V_R|$.

8 We compute average clustering by transforming directed graphs into undirected graphs, and we compute average shortest path lengths on the SCC.

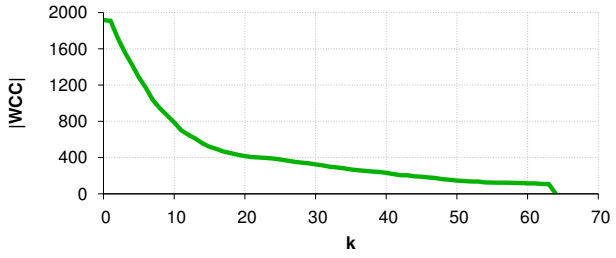


Fig. 5. k-core: size of the *Inclusion* graph WCC as nodes with degree $\leq k$ are recursively removed.

Betweenness Centrality	Weighted PageRank
google-analytics	doubleclick
doubleclick	googlesyndication
googleadservices	2mdn
facebook	adnxs
googletagmanager	google
googlesyndication	adsafeprotected
adnxs	google-analytics
google	scorecardresearch
addthis	krxd
criteo	rubiconproject

Table 2. Top 10 nodes ranked by betweenness centrality and weighted PageRank in the *Inclusion* graph.

network that have disassortative connectivity, which we examine in the next section.

4.2 Cores and Communities

We now examine how nodes in the *Inclusion* graph connect to each other using two metrics: k-cores and community detection. The k-core of a graph is the subset of a graph (nodes and edges) that remain after recursively removing all nodes with degree $\leq k$. By increasing k, the loosely connected periphery of a graph can be stripped away, leaving just the dense core. In our scenario, this corresponds to the high-degree ad exchanges, ad networks, and trackers that facilitate the connections between publishers and advertisers.

Figure 5 plots k versus the size of the WCC for the *Inclusion* graph. The plot shows that the core of the *Inclusion* graph rapidly declines in size as k increases, which highlights the interdependence between A&A domains and the lack of a distinct core.

Next, to examine the community structure of the *Inclusion* graph, we utilized three different community detection algorithms: label propagation by Raghavan et al. [64], Louvain modularity maximization [12], and the centrality-based Girvan–Newman [27] algorithm. We chose these algorithms because they attempt to find communities using fundamentally different approaches.

Unfortunately, after running these algorithms on the largest WCC, the results of our community analysis were negative. Label propagation clustered all nodes into a single community. Louvain found 14 communities with an overall modularity score of 0.44 (on a scale of -1 to 1 where 1 is entirely disjoint clusters). The largest community contains 771 nodes (40% of all nodes) and 3252 edges (12% of all edges). Out of 771 nodes, 37% are A&A. However, none of the 14 communities corresponded to meaningful groups of nodes, either segmented by type (e.g., publishers, SSPs, DSPs, etc.) or segmented by ad exchange (e.g., customers and partners centered around DoubleClick). This is a known deficiency in modularity maximization based methods, that they tend to produce communities with no real-world correspondence [5]. Girvan–Newman found 10 communities, with the largest community containing 1,097 nodes (57% of all nodes) and 16,424 edges (63% of all edges). Out of 1,097 nodes, 64% are A&A. However, the modularity score was zero, which means that the Girvan–Newman communities contain a random assortment of internal and external (cross-cluster) edges.

Overall, these results demonstrate that the web display ad ecosystem is not balkanized into distinct groups of companies and publishers that partner with each other. Instead, the ecosystem is highly interdependent, with no clear delineations between groups or types of A&A companies. This result is not surprising considering how dense the *Inclusion* graph is.

4.3 Node Importance

In this section, we focus on the importance of specific nodes in the *Inclusion* graph using two metrics: betweenness centrality and weighted PageRank. As before, we focus on the largest WCC. The betweenness centrality for a node n is defined as the fraction of all shortest paths on the graph that traverse n. In our scenario, nodes with high betweenness centrality represent the key pathways for tracking information and impressions to flow from publishers to the rest of the ad ecosystem. For weighted PageRank, we weight each edge in the *Inclusion* graph based on the number of times we observe it in our raw data. In essence, weighted PageRank identifies the nodes that receive the largest amounts of tracking data and impressions throughout each graph.

Table 2 shows the top 10 nodes in the *Inclusion* graph based on betweenness centrality and weighted PageRank. Prominent online advertising companies are well represented, including AppNexus (*adnxs*), Facebook, and Integral Ad Science (*adsafeprotected*). Similar to prior work, we find that Google's advertising domains (including DoubleClick and *2mdn*) are the most prominent overall [29]. Unsurprisingly, these companies all provide platforms, i.e., SSPs, ad exchanges, and ad networks. We also observe trackers like Google Analytics and Tag Manager. Interestingly, among 14 unique domains across the two lists, ten only appear in a single list. This suggests that the most important domains in terms of connectivity are not necessarily the ones that receive the highest volume of HTTP requests.

5 Information Diffusion

In § 4, we examined the descriptive characteristics of the *Inclusion* graph, and discuss the implications of this graph structure on our understanding of the online advertising ecosystem. In this section, we take the next step and present a concrete use case for the *Inclusion* graph: modeling the diffusion of user tracking data across the ad ecosystem under different types of ad and tracker blocking (e.g., AdBlock Plus and Ghostery). We model the flow of information across the *Inclusion* graph, taking into account different blocking strategies, as well as the design of RTB systems and empirically observed transition probabilities from our crawled dataset.

5.1 Simulation Goals

Simulation is an important tool for helping to understand the dynamics of the (otherwise opaque) online advertising industry. For example, Gill et al. used data-driven simulations to model the distribution of revenue amongst online display advertisers [26].

Here, we use simulations to examine the flow of browsing history data to trackers and advertisers. Specifically, we ask:

1. How many user impressions (i.e., page visits) to publishers can each A&A domain observe?
2. What fraction of the unique publishers that a user visits can each A&A domain observe?
3. How do different blocking strategies impact the number of impressions and fraction of publishers observed by each A&A domain?

These questions have direct implications for understanding users' online privacy. The first two questions are about quantifying a user's online footprint, i.e., how much of their browsing history can be recorded by different companies. In contrast, the third question investigates how well different blocking strategies perform at protecting users' privacy.

5.2 Simulation Setup

To answer these questions, we simulate the browsing behavior of typical users using the methodology from Burklen et al. [14].[9] In particular, we simulate a user browsing publishers over discreet time steps. At each time step our simulated user decides whether to remain on the current publisher according to a Pareto distribution (exponent = 2), in which case they generate a new impression on that publisher. Otherwise, the user browses to a new publisher, which is chosen based on a Zipf distribution over the Alexa ranks of the publishers. Burklen et al. developed this browsing model based on large-scale observational traces, and derive the distributions and their parameters empirically. This browsing model has been successfully used to drive simulated experiments in other work [40].

We generated browsing traces for 200 users. On average, each user generated 5,343 impressions on 190 unique publishers. The publishers are selected from the 888 unique first-party websites in our dataset (see § 3.1).

During each simulated time step the user generates an impression on a publisher, which is then forwarded to all A&A domains that are directly connected to the publisher. This emulates a webpage with multiple slots for display ads, each of which is serviced by a different SSP or ad exchange. However, it is insufficient to simply forward the impression to the A&A domains directly connected to each publisher; we also must account for ad exchanges and RTB auctions [10, 58], which may cause the impression to spread farther on the graph. We discuss this process next. The simulated time step ends when all impressions arrive at A&A domains that do not forward them. Once all outstanding impressions have terminated, time increments and our simulated user generates a new impression, either from their currently selected publisher or from a new publisher.

9 To the best of our knowledge, there are no other empirically validated browsing models besides [14].

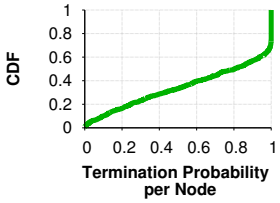


Fig. 6. CDF of the *termination probability* for A&A nodes.

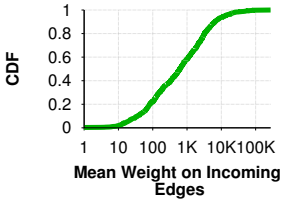


Fig. 7. CDF of the weights on incoming edges for A&A nodes.

5.2.1 Impression Propagation

Our simulations must account for *direct* and *indirect* propagation of impressions. Direct flows occur when one A&A domain sells or redirects an impression to another A&A domain. We refer to these flows as "direct" because they are observable by the web browser, and are thus recorded in our dataset. Indirect flows occur when an ad exchange solicits bids on an impression. The advertisers in the auction learn about the impression, but this is not directly observable to the browser; only the winner is ultimately known.

Direct Propagation. To account for direct propagation, we assign a *termination probability* to each A&A node in the *Inclusion* graph that determines how often it serves an ad itself, versus selling the impression to a partner (and redirecting the user's browser accordingly). We derive the termination probability for each A&A node empirically from our dataset. When an impression is sold, we determine which neighboring node purchases the impression based on the weights of the outgoing edges. For a node a_i, we define its set of outgoing neighbors as $\mathcal{N}_o(a_i)$. The probability of selling to neighbor $a_j \in \mathcal{N}_o(a_i)$ is $w(a_i \rightarrow a_j)/\sum_{\forall a_y \in \mathcal{N}_o(a_i)} w(a_i \rightarrow a_y)$, where $w(a_i \rightarrow a_j)$ is the weight of the given edge.

Figure 6 shows the *termination probability* for A&A nodes in the *Inclusion* graph. We see that 25% of the A&A nodes have a termination probability of one, meaning that they never sell impressions. The remaining 75% of A&A nodes exhibit a wide range of termination probabilities, corresponding to different business models and roles in the ad ecosystem. For example, DoubleClick, the most prominent ad exchange, has a termination probability of 0.35, whereas Criteo, a well-known advertiser specializing in retargeting, has a termination probability of 0.63.

Figure 7 shows the mean incoming edge weights for A&A nodes in the *Inclusion* graph. We observe that the distribution is highly skewed towards nodes with extremely high average incoming weights (note that the x-axis is in log scale). This demonstrates that heavy-hitters like DoubleClick, GoogleSyndication, OpenX, and Facebook are likely to purchase impressions that go up for auction in our simulations.

Indirect Propagation. Unfortunately, precisely accounting for indirect propagation is not currently possible, since it is not known exactly which A&A domains are ad exchanges, or which pairs of A&A domains share information. To compensate, we evaluate three different indirect impression propagation models:

- **Cookie Matching-Only:** As we note in § 3.2, the Bashir et al. [10] dataset includes 200 empirically validated pairs of A&A domains that match cookies. In this model, we treat these 200 edges as ground-truth and only indirectly disseminate impressions along these edges. Specifically, if a_i observes an impression, it will indirectly share with a_j iff $a_i \rightarrow a_j$ exists and is in the set of 200 known cookie matching edges. This is the most conservative model we evaluate, and it provides a lower-bound on impressions observed by A&A domains.
- **RTB Relaxed:** In this model, we assume that each A&A domain that observes an impression, indirectly shares it with all A&A domains that it is connected to. Although this is the correct behavior for ad exchanges like Rubicon and DoubleClick, it is not correct for every A&A domain. This is the most liberal model we evaluate, and it provides an upper-bound on impressions observed by A&A domains.
- **RTB Constrained:** In this model, we select a subset of A&A domains E to act as ad exchanges. Whenever an A&A domain in E observes an impression, it shares it with all directly connected A&A domains, i.e., to solicit bids. This model represents a more realistic view of information diffusion than the Cookie Matching-Only and RTB Relaxed models because the graph contains few but extremely well connected exchanges.

For RTB Constrained, we select all A&A nodes with out-degree ≥ 50 and in/out degree ratio r in the range $0.7 \leq r \leq 1.7$ to be in E. These thresholds were chosen after manually looking at the degrees and ratios for known ad exchanges and ad exchanges marked by Bashir et al. [10]. This results in $|E| = 36$ A&A nodes being chosen as ad exchanges (out of 1,032 total A&A domains in the *Inclusion* graph). We enforce restrictions on r because A&A nodes with disproportionately large amounts of incoming edges are likely to be trackers (in-

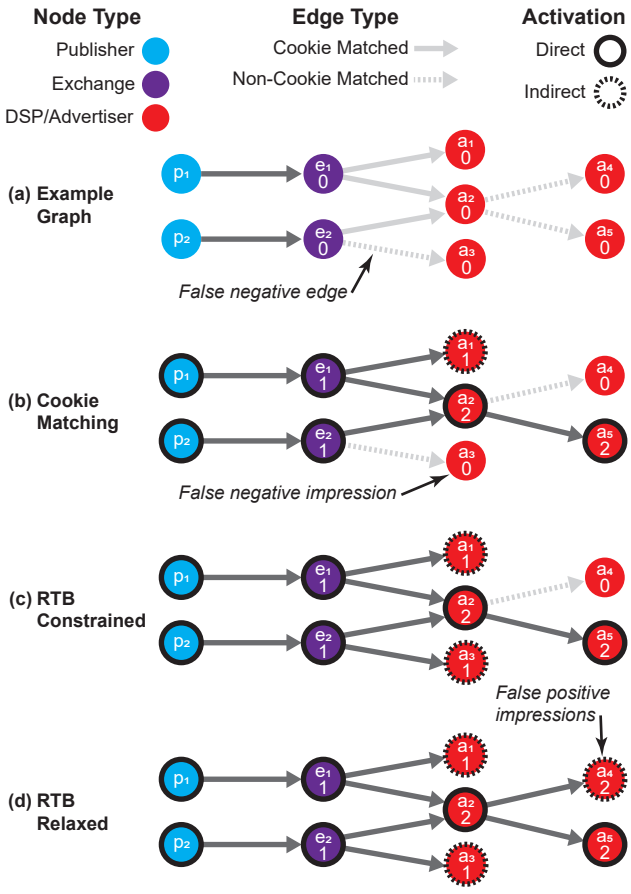


Fig. 8. Examples of our information diffusion simulations. The observed impression count for each A&A node is shown below its name. **(a)** shows an example graph with two publishers and two ad exchanges. Advertisers a_1 and a_3 participate in the RTB auctions, as well as DSP a_2 that bids on behalf of a_4 and a_5. **(b)–(d)** show the flow of data (dark grey arrows) when a user generates impressions on p_1 and p_2 under three diffusion models. In all three examples, a_2 purchases both impressions on behalf of a_5, thus they both *directly* receive information. Other advertisers *indirectly* receive information by participating in the auctions.

formation enters but is not forwarded out), while those with disproportionately large amounts of outgoing edges are likely SSPs (they have too few incoming edges to be an ad exchange). Table 6 in the appendix shows the domains in E, including major, known ad exchanges like App Nexus, Advertising.com, Casale Media, DoubleClick, Google Syndication, OpenX, Rubicon, Turn, and Yahoo. 150 of the 200 known cookie matching edges in our dataset are covered by this list of 36 nodes.

Figure 8 shows hypothetical examples of how impressions disseminate under our indirect models. Figure 8(a) presents the scenario: a graph with two publishers connected to two ad exchanges and five advertisers. a_2 is a bidder in both exchanges, and serves as a DSP for a_4 and a_5 (i.e., it services their ad campaigns by bidding on their behalf). Light grey edges capture cases where the two endpoints have been observed cookie matching in the ground-truth data. Edge $e_2 \rightarrow a_3$ is a false negative because matching has not been observed along this edge in the data, but a_3 must match with e_2 to meaningfully participate in the auction.

Figure 8(b)–(d) show the flow of impressions under our three models. In all three examples, a user visits publishers p_1 and p_2, generating two impressions. Further, in all three examples a_2 wins both auctions on behalf of a_5; thus e_1, e_2, a_2, and a_5 are guaranteed to observe impressions. As shown in the figure, a_2 and a_5 observe both impressions, but other nodes may observe zero or more impressions depending on their position and the dissemination model. In Figure 8(b), a_3 does not observe any impressions because its incoming edge has not been labeled as cookie matched; this is a false negative because a_3 participates in e_2's auction. Conversely, in Figure 8(d), all nodes always share all impressions, thus a_4 observes both impressions. However, these are false positives, since DSPs like a_2 do not routinely share information amongst all their clients.

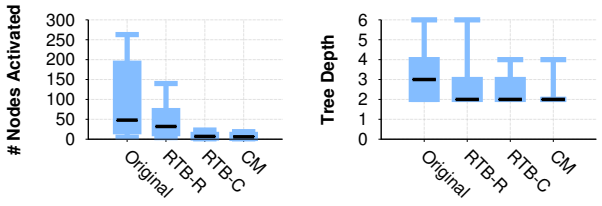
5.2.2 Node Blocking

To answer our third question, we must simulate the effect of "blocking" A&A domains on the *Inclusion* graph. A simulated user that blocks A&A domain a_j will not make direct connections to it (the solid outlines in Figure 8). However, blocking a_j does **not** prevent a_j from tracking users indirectly: if the simulated user contacts ad exchange a_i, the impression may be forwarded to a_j during the bidding process (the dashed outlines in Figure 8). For example, an extension that blocks a_2 in Figure 8 will prevent the user from seeing an ad, as well as prevent information flow to a_4 and a_5. However, blocking a_2 does not stop information from flowing to e_1, e_2, a_1, a_3, and even a_2!

We evaluate five different blocking strategies to compare their relative impact on user privacy under our three impression propagation models:

1. We randomly blocked 30% (310) of the A&A nodes from the *Inclusion* graph.[10]
2. We blocked the top 10% (103) of A&A nodes from the *Inclusion* graph, sorted by weighted PageRank.

10 We also randomly blocked 10% and 20% of A&A nodes, but the simulation results were very similar to that of random 30%.



(a) Number of nodes **(b)** Tree depth

Fig. 9. Comparison of the original and simulated inclusion trees. Each bar shows the 5^{th}, 25^{th}, 50^{th} (in black), 75^{th}, and 95^{th} percentile value.

3. We blocked all 594 A&A nodes from the Ghostery [25] blacklist.
4. We blocked all 412 A&A nodes from the Disconnect [18] blacklist.
5. We emulated the behavior of AdBlock Plus [2], which is a combination of whitelisting A&A nodes from the Acceptable Ads program [73], and blacklisting A&A nodes from EasyList [19]. After whitelisting, 634 A&A nodes are blocked.

We chose these methods to explore a range of graph theoretic and practical blocking strategies. Prior work has shown that the global connectivity of small-world graphs is resilient against random node removal [13], but we would like to empirically determine if this is true for ad network graphs as well. In contrast, prior work also shows that removing even a small fraction of top nodes from small-world graphs causes the graph to fracture into many subgraphs [50, 74]. Ghostery and Disconnect are two of the most widely-installed tracker blocking browser extensions, so evaluating their blacklists allows us to quantify how good they are at protecting users' privacy. Finally, AdBlock Plus is the most popular ad blocking extension [45, 62], but contrary to its name, by default it whitelists A&A companies that pay to be part of its Acceptable Ads program [3]. Thus, we seek to understand how effective AdBlock Plus is at protecting user privacy under its default behavior.

5.3 Validation

To confirm that our simulations are representative of our ground-truth data, we perform some sanity checks. We simulate a single user in each model (who generates 5K impressions) and compare the resulting simulated inclusion trees to the original, real inclusion trees.

First, we look at the number of nodes that are activated by direct propagation in trees rooted at each publisher. Figure 9a shows that our models are conservative in that they generate smaller trees: the median original tree contains 48 nodes, versus 32, seven, and six from our models. One caveat to this is that publishers in our simulated trees have a wider range of fan-outs than in the original trees. The median publishers in the original and simulated trees have 11 and 12 neighbors, respectively, but the 75^{th} percentile trees have 16 and 30 neighbors, respectively.

Second, we investigate the depth of the inclusion trees. As shown in Figure 9b, the median tree depth in the original trees is three, versus two in all our models. The 75^{th} percentile tree depth in the original data is four, versus three in the RTB Relaxed and RTB Constrained models, and two in the most restrictive Cookie Matching-Only model. These results show that overall, our models are conservative in that they tend to generate slightly shorter inclusion trees than reality.

Third, we look at the set of A&A domains that are included in trees rooted at each publisher. For a publisher p that contacts a set A_p^o of A&A domains in our **o**riginal data, we calculate $f_p = |A_p^s \cap A_p^o| / |A_p^o|$, where A_p^s is the set of A&A domains contacted by p in **s**imulation. Figure 10 plots the CDF of f_p values for all publishers in our dataset, under our three models. We observe that for almost 80% publishers, 90% A&A domains contacted in the original trees are also contacted in trees generated by the RTB Relaxed model. This falls to 60% and 16% as the models become more restrictive.

Fourth, we examine the number of ad exchanges that appear in the original and simulated trees. Examining the ad exchanges is critical, since they are responsible for all indirect dissemination of impressions. As shown in Figure 11, inclusion trees from our simulations contain an order of magnitude fewer ad exchanges than the original inclusion trees, regardless of model.[11] This suggests that indirect dissemination of impressions in our models will be conservative relative to reality.

Number of Selected Exchanges. Finally, we investigate the impact of exchanges in the RTB Constrained model. We select the top x A&A domains by out-degree to act as exchanges (subject to their in/out degree ratio r being in the range $0.7 \leq r \leq 1.7$), then execute a simulation. As shown in Figure 12, with 20

11 Because each of our models assumes that a different set of A&A nodes are ad exchanges, we must perform three corresponding counts of ad exchanges in our original trees.

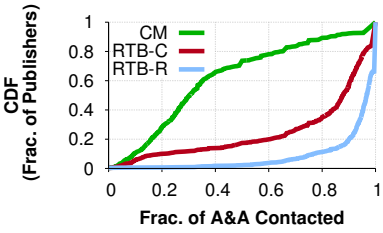


Fig. 10. CDF of the fractions of A&A domains contacted by publishers in our original data that were **also** contacted in our three simulated models.

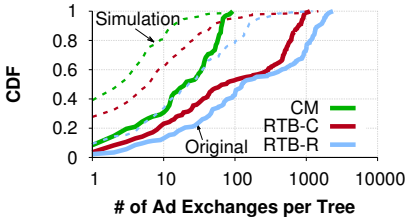


Fig. 11. Number of ad exchanges in our original (solids lines) and simulated (dashed lines) inclusion trees.

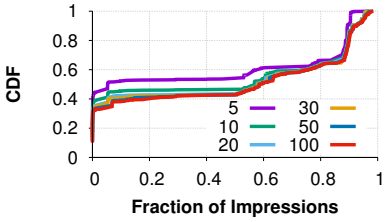


Fig. 12. Fraction of impressions observed by A&A domains in RTB-C model when top x exchanges are selected.

Blocking Scenarios	Cookie Matching-Only %E	Cookie Matching-Only %W	RTB Constrained %E	RTB Constrained %W	RTB Relaxed %E	RTB Relaxed %W
No Blocking	16.9	31.0	33.9	55.9	71.8	81.3
AdBlock Plus	12.3	28.0	25.6	50.3	48.4	68.6
Random 30%	12.1	21.8	22.1	34.2	48.7	54.8
Ghostery	3.52	9.87	6.82	18.2	13.5	21.9
Top 10%	6.03	5.01	8.18	5.52	26.8	13.4
Disconnect	2.98	3.66	4.72	6.01	16.3	11.6

Table 3. Percentage of **E**dges that are triggered in the *Inclusion* graph during our simulations under different propagation models and blocking scenarios. We also show the percentage of edge **W**eights covered via triggered edges.

Cookie Matching-Only		RTB Constrained		RTB Relaxed	
doubleclick	90.1	google-analytics	97.1	pinterest	99.1
criteo	89.6	quantserve	92.0	doubleclick	99.1
quantserve	89.5	scorecardresearch	91.9	twitter	99.1
googlesyndication	89.0	youtube	91.8	googlesyndication	99.0
flashtalking	88.8	skimresources	91.6	scorecardresearch	99.0
mediaforge	88.8	twitter	91.3	moatads	99.0
adsrvr	88.6	pinterest	91.2	quantserve	99.0
dotomi	88.6	criteo	91.2	doubleverify	99.0
steelhousemedia	88.6	addthis	91.1	crwdcntrl	99.0
adroll	88.6	bluekai	91.1	adsrvr	99.0

Table 4. Top 10 nodes that observed the most impressions under our simulations with no blocking.

or more exchanges the distribution of impressions observed by A&A domains stops growing, i.e., our RTB Constrained model is relatively insensitive to the number of exchanges. This is not surprising, given how dense the *Inclusion* graph is (see § 4). We observed similar results when we picked top nodes based on PageRank.

5.4 Results

We take our 200 simulated users and "play back" their browsing traces over the unmodified *Inclusion* graph, as well as graphs where nodes have been blocked using the strategies outlined above. We record the total number of impressions observed by each A&A domain, as well as the fraction of unique publishers observed by each A&A domain under different impression propagation models.

Triggered Edges. Table 3 shows the percentage of edges between A&A nodes that are triggered in the *Inclusion* graph under different combinations of impression propagation models and blocking strategies. No blocking/RTB Relaxed is the most permissive case; all other cases have less edges and weight because (1) the propagation model prevents specific A&A edges from being activated and/or (2) the blocking scenario explicitly removes nodes. Interestingly, AdBlock Plus fails to have significant impact relative to the No Blocking baseline, in terms of removing edges or weight, under the Cookie Matching-Only and RTB Constrained models. Further, the top 10% blocking strategy removes less edges than Disconnect or Ghostery, but it reduces the remaining edge weight to roughly the same level as Disconnect, whereas Ghostery leaves more high-weight edges intact. These observations help to explain the outcomes of our simulations, which we discuss next.

No Blocking. First, we discuss the case where no A&A nodes are blocked in the graph. Figure 13 shows the fraction of total impressions (out of ∼5,300) and fraction of unique publishers (out of ∼190) observed by A&A domains under different propagation models. We find that the distribution of observed impressions under RTB Constrained is very similar to that of RTB Relaxed, whereas observed impressions drop dramatically under Cookie Matching-Only model. Specifically, the top 10% of A&A nodes in the *Inclusion* graph (sorted by impression count) observe more than 97% of the impressions in RTB Relaxed, 90% in RTB Constrained, and 29% in Cookie Matching-Only. We observe similar patterns for fractions of publishers observed across the three indirect propogating models. Recall that the Cookie Matching-Only and RTB Relaxed models function as lower- and upper-bounds on observability; that

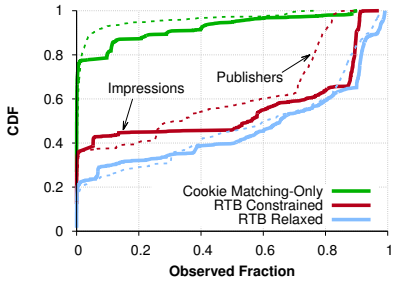
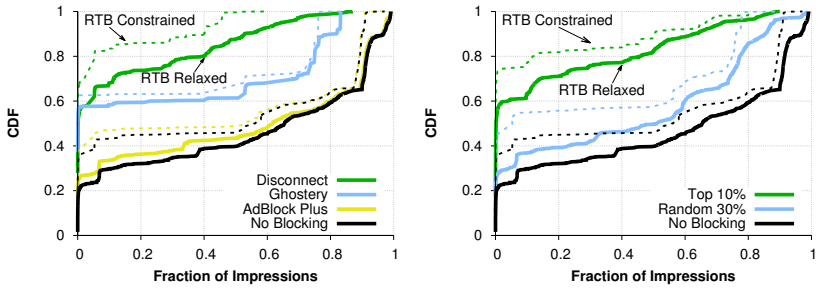


Fig. 13. Fraction of impressions (solid lines) and publishers (dashed lines) observed by A&A domains under our three models, without any blocking.



(a) Disconnect, Ghostery, AdBlock Plus **(b)** Top 10% and Random 30% of nodes

Fig. 14. Fraction of impressions observed by A&A domains under the **RTB Constrained** (dashed lines) and **RTB Relaxed** (solid lines) models, with various blocking strategies.

the results from the RTB Constrained model are so similar to the RTB Relaxed model is striking, given that only 36 nodes in the former spread impressions indirectly, versus 1,032 in the latter.

Although the overall fraction of observed impressions drops significantly in the Cookie Matching-Only model, Table 4 shows that the top 10 A&A domains observe 99%, 96%, and 89% of impressions on average under RTB Relaxed, RTB Constrained, and Cookie Matching-Only respectively. Some of the top ranked nodes are expected, like DoubleClick, but other cases are more interesting. For example, Pinterest is connected to 178 publishers and 99 other A&A domains. In the Cookie Matching-Only model, it ranks 47 because it is directly embedded in relatively few publishers, but it ascends up to rank seven and one, respectively, once indirect sharing is accounted for. This drives home the point that although Google is the most pervasively embedded advertiser around the web [15, 65], there are a roughly 52 other A&A companies that also observe greater than 91% of users' browsing behaviors (in the RTB Constrained model), due to their participation in major ad exchanges.

With Blocking. Next, we discuss the results when AdBlock Plus (i.e., the Acceptable Ads whitelist and EasyList blacklist) is used to block nodes. AdBlock Plus has essentially zero impact on the fraction of impressions observed by A&A domains: the results in Figure 14a under the RTB Constrained and RTB Relaxed models are almost coincident with those for the models when no blocking is applied at all. The problem is that the major ad networks and exchanges are all present in the Acceptable Ads whitelist, and thus all of their partners are also able to observe the impressions, even if they are (sometimes) prevented from actually showing ads to the user. Indeed, the top 10 nodes in Table 4 with no blocking and in Table 5 with AdBlock Plus are almost identical, save for some reordering.

Next, we examine Ghostery and Disconnect in Figure 14a. As expected, the amount of information seen by A&A domains decreases when we block domains from these blacklists. Disconnect's blacklist does a much better job of protecting users' privacy in our simulations: after blocking nodes using the Disconnect blacklist, 90% of the nodes see less than 40% of the impressions in the RTB Constrained model, and less than 53% in the RTB Relaxed model. In contrast, when using the Ghostery blacklist, 90% of the nodes see less than 75% of the impressions in both RTB models. Table 5 shows that top 10 A&A domains are only able to observe at most 40–59% and 73–83% of impressions when the Disconnect and Ghostery blacklists are used, respectively, depending on the indirect propagation model.

As shown in Figure 14b, blocking the top 10% of A&A nodes from the *Inclusion* graph (sorted by weighted PageRank) causes almost as much reduction in observed impressions as Disconnect. Table 5 helps to orient the top 10% blocking strategy versus Disconnect and Ghostery in terms of overall reduction in impression observability and the impact on specific A&A domains. In contrast, blocking 30% of the A&A nodes at random has more impact than AdBlock Plus, but less than Disconnect and Ghostery. Top 10 nodes under the "no blocking" and "random 30%" (not shown) strategies observe similar impression fractions. Both of these results agree with the theoretical expectations for small-world graphs, i.e., their connectivity is resilient against random blocking, but not necessarily targeted blocking.

We do not show results for our most restrictive model (i.e., Cookie Matching-Only) in Figure 14, since the majority of A&A companies view almost zero impressions. Specifically, 90% of A&A companies view less

AdBlock Plus				Disconnect				Ghostery				Top 10 %			
CM-Only	%	RTB Constrained	%	CM-Only	%	RTB Constrained	%	CM-Only	%	RTB Constrained	%	CM-Only	%	RTB Constrained	%
doubleclick	90.0	google-analytics	97.0	amazonaws	43.7	amazonaws	59.3	criteo	75.0	google-analytics	83.1	rubiconproject	64.3	doubleclick	80.6
quantserve	89.5	youtube	91.7	3lift	41.5	revenuemantra	51.6	googlesyndication	74.7	youtube	77.4	amazon-adsystem	64.2	doubleverify	80.6
criteo	89.4	quantserve	91.6	zergnet	40.9	bidswitch	50.8	2mdn	74.5	betrad	76.2	googlesyndication	64.2	googlesyndication	80.6
googlesyndication	88.9	scorecardresearch	91.6	celtra	40.5	jwpltx	50.5	doubleclick	74.5	acexedge	76.2	mathtag	52.5	moatads	80.6
dotomi	88.6	skimresources	91.3	sonobi	40.4	basebanner	50.4	adnxs	73.3	vindicosuite	76.2	undertone	52.1	2mdn	80.6
flashtalking	88.6	twitter	91.1	bzgint	40.2	zergnet	46.0	adroll	73.3	2mdn	76.1	sitescout	50.1	twitter	80.6
adroll	88.5	pinterest	91.0	eyeviewads	40.2	sonobi	45.8	adsrvr	73.3	360yield	76.1	doubleclick	49.8	bluekai	80.6
adsrvr	88.5	addthis	90.9	simplereach	40.0	adnxs	45.8	adtechus	73.3	adadvisor	76.1	adtech	49.7	google-analytics	80.5
mediaforge	88.5	criteo	90.9	richmetrics	39.9	adsafeprotected	45.8	advertising	73.3	adap	76.1	adnxs	49.7	media	80.5
steelhousemedia	88.5	bluekai	90.8	kompasads	39.9	adsrvr	45.8	amazon-adsystem	73.3	adform	76.1	mediaforge	49.6	exelator	80.5

Table 5. Top 10 nodes that observed the most impressions in the **Cookie Matching-Only** and **RTB Constrained** models under various blocking scenarios. The numbers for the **RTB Relaxed** model (not shown) are slightly higher than those for RTB Constrained. Results under blocking random 30% nodes (not shown) are slighlty lower than no blocking.

than 0.2%, 0.3%, and 11% of the impressions under Ghostery, Disconnect, and top 10% blocking. However, we do present the number of impressions seen by top 10 A&A domains in the Cookie Matching-Only model in Table 5, which shows that even under strict blocking strategies, top advertising companies still view 40–75% of the impressions.

Summary. Overall, there are three takeaways from our simulations. *First*, the "no blocking" simulation results show that top A&A domains are able to see the vast majority of users' browsing history, which is extremely troubling from a privacy perspective. For example, even under the most constrained propagation model (Cookie Matching-Only), DoubleClick still observes 90% of all impressions generated by our simulated users. *Second*, it is troubling to observe that AdBlock Plus barely improves users' privacy, due to the Acceptable Ads whitelist containing high-degree ad exchanges. *Third*, we find that users can improve their privacy by blocking A&A domains, but that the choice of blocking strategy is critically important. We find that the Disconnect blacklist offers the greatest reduction in observable impressions, while Ghostery offers significantly less protection. However, even when strong blocking is used, top A&A domains still observe anywhere from 40–80% of simulated users' impressions.

5.5 Random Browsing Model

Thus far, we have analyzed results for users that follow the browsing model from Burklen et al. [14]. This is, to the best of our knowledge, the only empirically validated browsing model.

To check the consistency of our simulation results, we ran additional simulations using a random browsing model, where the user chooses publishers purely at random, and chooses whether to remain on a publisher or depart using a coin flip.

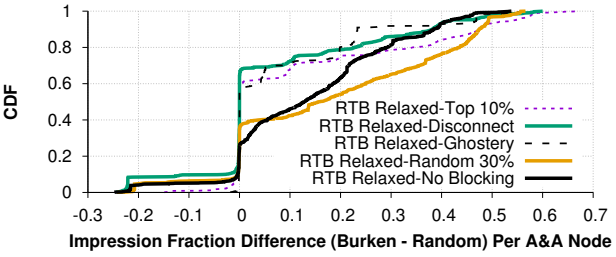


Fig. 15. Difference of impression fractions observed by A&A nodes with simulations between Burklen et al. [14] and the random browsing model.

We plot the results of the random simulations in Figure 15 as the difference in fraction of impressions observed by A&A domains under the RTB Relaxed model. Zero indicates that an A&A domain observed the same fraction of impressions in both the Burklen et al. and random user simulations, while <0 (>0) indicates that the node observed more impressions in the random (Burklen et al.) simulations. Between 20–60% of A&A nodes observe the same amount of impressions regardless of model, but this is because these nodes all observe **zero** impressions (i.e., they are blocked). This is why the fraction of A&A nodes that do not change between the browsing models is greatest with Disconnect. Although up to 10% of A&A nodes observe more impressions under the random browsing model, the majority of A&A nodes that observe at least one impression observe more overall under the Burklen et al. model.

Overall, Figure 15 demonstrates that the baseline browsing behavior exhibited by a user does have a significant impact on their visibility to A&A companies. For example, using the Burklen et al. model [14], the selected publishers contact top 10 A&A domains (sorted by PageRank) 2.6× more than those selected by the random browsing model (and 4.6× if we consider the top 10 A&A domains sorted by betweenness centrality).

Importantly, however, the relative effectiveness of blocking strategies remains the same under a random

browsing model. Disconnect still performed the best, followed by top 10%, Ghostery, random 30%, and then AdBlock Plus. This suggests that our findings with respect to the efficacy of blocking strategies generalizes to users with different browsing behaviors.

6 Limitations

As with all simulated models, there are some limitations to our work.

First, our models of indirect impression dissemination are approximations. The Cookie Matching-Only and RTB Relaxed models should be viewed as lower- and upper-estimates, respectively, on the dissemination of impressions, not as accurate reflections of reality (for the reasons highlighted in Figure 8). We believe that the RTB Constrained model is a reasonable approximation, but even it has flaws: it may still exhibit false positives, if non-exchanges are included in the set of exchanges E, and false negatives if an actual exchange is not included in E. Furthermore, it is not clear in general if ad exchanges always forward all impressions to all partners. For example, *private exchanges* that connect high-value publishers (e.g., The New York Times) to select pools of advertisers behave differently than their public cousins.

Second, our results are dependent on assumptions about the browsing behavior of users. We present results from two browsing models in § 5.5 and show that many of our headline results are robust. However, these findings should not be over-generalized: they are representative for an average user, yet specific individuals may experience different amounts of tracking.

Third, we must translate rules from the EasyList blacklist and the Acceptable Ads whitelist to use them in our simulations. Both of these lists include rules containing regular expressions, URLs, and even snippets of CSS; we simplify them to lists of effective 2^{nd}-level domains. Due to this translation, we may over-estimate impressions seen by the whitelisted A&A domains, and under-estimate impressions seen by blacklisted A&A domains. Note that the Ghostery and Disconnect blacklists are not affected by these issues.

Fourth, we analyze a dataset that was collected in December 2015. The structure of the *Inclusion* graph has almost certainly changed since then. Furthermore, the edge weights between nodes may differ depending on the initial set of publishers that are crawled. Although we demonstrate in § 5.3 that our dataset covers the vast majority of A&A domains, the connectivity and weights between A&A domains may change over time.

Fifth, our dataset does not cover the mobile advertising ecosystem, which is known to differ from the web ecosystem [72]. Thus our results likely do not generalize to this area.

7 Conclusion

In this paper, we introduce a novel graph model of the advertising ecosystem called an *Inclusion* graph. This representation is enabled by advances in browser instrumentation [6, 41] that allow researchers to capture the precise inclusion relationships between resources from different A&A domains [10]. Using a large, crawled dataset from [10], we show that the ad ecosystem is extremely dense. Furthermore, we compare our *Inclusion* graph representation to a *Referer* graph representation proposed by prior work [29], and show that the *Referer* graph has substantive structural differences that are caused by erroneously attributed edges.

We show that our *Inclusion* graph can be used to implement empirically-driven simulations of the online ad ecosystem. Our results demonstrate that under a variety of assumptions about user browsing and advertiser interaction behavior, top A&A companies observe the vast majority of users' browsing history. Even under realistic conditions where only a small number of well-connected ad exchanges indirectly share impressions, 10% of A&A companies observe more than 90% impressions and 82% publishers.

We also evaluate a variety of ad and tracker blocking strategies in the context of our models, to understand their effectiveness at stopping A&A companies from learning users' browsing history. On one hand, we find that blocking the top 10% of A&A domains, as well as the Disconnect blacklist, do significantly reduce the observation of users' browsing. On the other hand, even these strategies still leak 40–80% of users' browsing history to top A&A domains, under realistic assumptions. This suggests that users who truly care about privacy on the web should adopt the most stringent blocking tools available, such as EasyList and EasyPrivacy, or consider disabling JavaScript by default with an extension like uMatrix [28].

Acknowledgments

We thank all of the reviewers and our shepherd for their helpful feedback. This research was supported in part by NSF grants IIS-1408345 and IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proc. of CCS*, 2014.

[2] Adblock plus: Surf the web without annoying ads! eyeo GmbH. https://adblockplus.org.

[3] Allowing acceptable ads in adblock plus. eyeo GmbH. https://adblockplus.org/acceptable-ads.

[4] Alexa. The top 500 sites on the web. https://www.alexa.com/topsites/category/Top.

[5] Hélio Almeida, Dorgival Guedes, Wagner Meira, and Mohammed J. Zaki. Is there a best quality metric for graph clusters? In *Proc. of ECML PKDD*, 2011.

[6] Sajjad Arshad, Amin Kharraz, and William Robertson. Include me out: In-browser detection of malicious third-party content inclusions. In *Proc. of Intl. Conf. on Financial Cryptography*, 2016.

[7] Mika Ayenson, Dietrich James Wambach, Ashkan Soltani, Nathan Good, and Chris Jay Hoofnagle. Flash cookies and privacy ii: Now with html5 and etag respawning. *Available at SSRN 1898390*, 2011.

[8] Rebecca Balebako, Pedro G. Leon, Richard Shay, Blase Ur, Yang Wang, and Lorrie Faith Cranor. Measuring the effectiveness of privacy tools for limiting behavioral advertising. In *Proc. of W2SP*, 2012.

[9] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma, and S. Muthukrishnan. Adscape: Harvesting and analyzing online display ads. In *Proc. of WWW*, 2014.

[10] Muhammad Ahmad Bashir, Sajjad Arshad, , William Robertson, and Christo Wilson. Tracing information flows between ad exchanges using retargeted ads. In *Proc. of USENIX Security Symposium*, 2016.

[11] Muhammad Ahmad Bashir, Sajjad Arshad, and Christo Wilson. Recommended For You: A First Look at Content Recommendation Networks. In *Proc. of IMC*, 2016.

[12] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.

[13] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: Experiments and models. In *Proc. of WWW*, 2000.

[14] Susanne Burklen, Pedro Jose Marron, Serena Fritsch, and Kurt Rothermel. User centric walk: An integrated approach for modeling the browsing behavior of users on the web. In *Annual Symposium on Simulation*, April 2005.

[15] Aaron Cahn, Scott Alfeld, Paul Barford, and S. Muthukrishnan. An empirical study of web cookies. In *Proc. of WWW*, 2016.

[16] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. I always feel like somebody's watching me: Measuring online behavioural advertising. In *Proc. of ACM CoNEXT*, 2015.

[17] Big Commerce. Understanding Impressions in digital marketing. BigCommerce Inc., March 2016. https://www.bigcommerce.com/ecommerce-answers/impressions-digital-marketing/.

[18] Disconnect defends the digital you. Disconnect Inc. https://disconnect.me/.

[19] Easylist. The EasyList authors. https://easylist.to.

[20] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proc. of CCS*, 2016.

[21] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. Cookies that give you away: The surveillance implications of web tracking. In *Proc. of WWW*, 2015.

[22] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In *Proc of. Traffic Monitoring and Analysis*, 2014.

[23] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Tracking personal identifiers across the web. In *Proc. of PAM*, 2016.

[24] Arpita Ghosh, Mohammad Mahdian, Preston McAfee, and Sergei Vassilvitskii. To match or not to match: Economics of cookie matching in online advertising. In *Proc. of EC*, 2012.

[25] Ghostery: faster, cleaner, and safer browsing. Cliqz International GmbH i.Gr. https://www.ghostery.com/.

[26] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *Proc. of IMC*, 2013.

[27] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[28] GitHub. umatrix: Point and click matrix to filter net requests according to source, destination and type., October 2014. https://github.com/gorhill/uMatrix.

[29] R. Gomer, E. M. Rodrigues, N. Milic-Frayling, and M. C. Schraefel. Network analysis of third party tracking: User exposure to tracking cookies through search. In *Prof. of IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2013.

[30] Saikat Guha, Bin Cheng, and Paul Francis. Challenges in measuring online advertising systems. In *Proc. of IMC*, 2010.

[31] Mark D. Humphries and Kevin Gurney. Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLoS One*, 3(4), 2008.

[32] Muhammad Ikram, Hassan Jameel Asghar, Mohamed Ali Kâafar, Balachander Krishnamurthy, and Anirban Mahanti. Towards seamless tracking-free web: Improved detection of trackers via one-class learning. *PoPETs*, 2017(1):79–99, 2017.

[33] Sakshi Jain, Mobin Javed, and Vern Paxson. Towards mining latent client identifiers from network traffic. *PoPETs*, 2016(2):100–114, 2016.

[34] Vasiliki Kalavri, Jeremy Blackburn, Matteo Varvello, and Konstantina Papagiannaki. Like a pack of wolves: Community structure of web trackers. In *Proc. of Passive and Active Measurement*, 2016.

[35] Samy Kamkar. Evercookie - virtually irrevocable persistent cookies., September 2010. http://samy.pl/evercookie/.

[36] T. Kohno, A. Broido, and K. Claffy. Remote physical device fingerprinting. *IEEE Transactions on Dependable and Secure Computing*, 2(2):93–108, 2005.

[37] Balachander Krishnamurthy, Delfina Malandrino, and Craig E. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. In *Proc. of the Workshop on Usable Security*, 2007.

[38] Balachander Krishnamurthy, Konstantin Naryshkin, and Craig Wills. Privacy diffusion on the web: A longitudinal perspective. In *Proc. of WWW*, 2009.

[39] Balachander Krishnamurthy and Craig Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Proc. of W2SP*, 2011.

[40] James Larisch, David Choffnes, Dave Levin, Bruce M. Maggs, Alan Mislove, and Christo Wilson. CRLite: a Scalable System for Pushing all TLS Revocations to All Browsers. In *Proc. of IEEE Symposium on Security and Privacy*, 2017.

[41] Tobias Lauinger, Abdelberi Chaabane, Sajjad Arshad, William Robertson, Christo Wilson, and Engin Kirda. Thou shalt not depend on me: Analysing the use of outdated javascript libraries on the web. In *Proc of NDSS*, 2017.

[42] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. What matters to users?: Factors that affect users' willingness to share information with online advertisers. In *Proc. of the Workshop on Usable Security*, 2013.

[43] Tai-Ching Li, Huy Hang, Michalis Faloutsos, and Petros Efstathopoulos. Trackadvisor: Taking back browsing privacy from third-party trackers. In *Proc. of PAM*, 2015.

[44] Bin Liu, Anmol Sheth, Udi Weinsberg, Jaideep Chandrashekar, and Ramesh Govindan. Adreveal: Improving transparency into online targeted advertising. In *Proc. of HotNets*, 2013.

[45] Matthew Malloy, Mark McNamara, Aaron Cahn, and Paul Barford. Ad blockers: Global prevalence and impact. In *Proc. of IMC*, 2016.

[46] Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *Proc. of IEEE Symposium on Security and Privacy*, 2012.

[47] Aleecia M. McDonald and Lorrie Faith Cranor. Americans' attitudes about internet behavioral advertising practices. In *Proc. of WPES*, 2010.

[48] William Melicher, Mahmood Sharif, Joshua Tan, Lujo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. (do not) track me sometimes: Users' contextual preferences for web tracking. *PoPETs*, 2016(2):135–154, 2016.

[49] Georg Merzdovnik, Markus Huber, Damjan Buhov, Nick Nikiforakis, Sebastian Neuner, Martin Schmiedecker, and Edgar R. Weippl. Block me if you can: A large-scale study of tracker-blocking tools. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, 2017.

[50] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc. of IMC*, 2007.

[51] Keaton Mowery and Hovav Shacham. Pixel perfect: Fingerprinting canvas in html5. In *Proc. of W2SP*, 2012.

[52] Mozilla. Same-origin policy., May 2008. https://developer.mozilla.org/en-US/docs/Web/Security/Same-origin_policy.

[53] Muhammad Haris Mughees, Zhiyun Qian, and Zubair Shafiq. Detecting anti ad-blockers in the wild. *PoPETs*, 2017(3):130, 2017.

[54] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. Privaricator: Deceiving fingerprinters with little white lies. In *Proc. of WWW*, 2015.

[55] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proc. of IEEE Symposium on Security and Privacy*, 2013.

[56] Rishab Nithyanand, Sheharbano Khattak, Mobin Javed, Narseo Vallina-Rodriguez, Marjan Falahrastegar, Julia E. Powles, Emiliano De Cristofaro, Hamed Haddadi, and Steven J. Murdoch. Adblocking and counter blocking: A slice of the arms race. In *Proc. of FOCI*, 2016.

[57] Lukasz Olejnik, Claude Castelluccia, and Artur Janc. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. In *Proc. of HotPETs*, 2012.

[58] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. Selling off privacy at auction. In *Proc of NDSS*, 2014.

[59] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez, and Nikolaos Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay for your personal data? In *Proc. of IMC*, 2017.

[60] Fotios Papaodyssefs, Costas Iordanou, Jeremy Blackburn, Nikolaos Laoutaris, and Konstantina Papagiannaki. Web identity translator: Behavioral advertising and identity privacy with wit. In *Proc. of HotNets*, 2015.

[61] Tim Peterson. Facebook's liverail exits the ad server business, January 2016. http://adage.com/article/digital/facebook-s-liverail-exits-ad-server-business/302017/.

[62] Enric Pujol, Oliver Hohlfeld, and Anja Feldmann. Annoyed users: Ads and ad-block usage in the wild. In *Proc. of IMC*, 2015.

[63] PwC. Iab internet advertising revenue report, 2017 full year results. IAB, 2018. https://www.iab.com/wp-content/uploads/2018/05/IAB-2017-Full-Year-Internet-Advertising-Revenue-Report.REV2_.pdf.

[64] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76, Sep 2007.

[65] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *Proc. of NSDI*, 2012.

[66] Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie F. Cranor. Watching them watching me: Browser extensions impact on user privacy awareness and concern. In *Proc. of the Workshop on Usable Security*, 2016.

[67] Mike Shields. Facebook buys online video tech firm liverail, looks for bigger role in digital ads, July 2014. https://blogs.wsj.com/cmo/2014/07/02/facebook-buys-online-video-tech-firm-liverail-looks-for-bigger-role-in-digital-ads/.

[68] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. Flash cookies and privacy. In *AAAI Spring Symposium: Intelligent Information Privacy Management*, 2010.

[69] Oleksii Starov and Nick Nikiforakis. Extended tracking powers: Measuring the privacy diffusion enabled by browser extensions. In *Proc. of WWW*, 2017.

[70] Joseph Turow, Michael Hennessy, and Nora Draper. The tradeoff fallacy: How marketers are misrepresenting american consumers and opening them up to exploitation. Report from the Annenberg School for Communication, June 2015. https://www.asc.upenn.edu/sites/default/files/TradeoffFallacy_1.pdf.

[71] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In *Proc. of the Workshop on Usable Security*, 2012.

[72] Narseo Vallina-Rodriguez, Jay Shah, Alessandro Finamore, Yan Grunenberger, Konstantina Papagiannaki, Hamed Haddadi, and Jon Crowcroft. Breaking for commercials: Characterizing mobile advertising. In *Proc. of IMC*, 2012.

[73] Robert J. Walls, Eric D. Kilmer, Nathaniel Lageman, and Patrick D. McDaniel. Measuring the impact and perception of acceptable advertisements. In *Proc. of IMC*, 2015.

[74] Christo Wilson, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. Beyond Social Graphs: User Interactions in Online Social Networks and their Implications. *ACM Transactions on the Web (TWEB)*, 6(4):5:1–5:31, November 2012.

[75] Apostolis Zarras, Alexandros Kapravelos, Gianluca Stringhini, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. The dark alleys of madison avenue: Understanding malicious advertisements. In *Proc. of IMC*, 2014.

Node	Out Degree	In/Out Ratio
doubleclick	398	1.67
googleadservices	380	1.00
googlesyndication	318	1.28
adnxs	293	0.98
googletagmanager	253	0.98
2mdn	223	0.97
adsafeprotected	202	1.30
rubiconproject	191	1.14
mathtag	182	1.09
openx	170	0.79
pubmatic	157	0.96
casalemedia	136	1.10
krxd	134	1.08
adtechus	130	0.96
yahoo	124	1.31
chartbeat	124	0.96
contextweb	117	0.88
crwdcntrl	105	1.36
rlcdn	98	1.50
turn	86	1.48
amazon-adsystem	84	1.43
bzgint	72	0.86
monetate	72	0.76
rhythmxchange	71	1.13
rfihub	70	1.46
gigya	69	0.78
revsci	67	1.00
media	57	1.07
adtech	57	0.93
simplereach	57	0.84
tribalfusion	55	0.75
disqus	55	0.95
w55c	55	1.55
afy11	54	1.33
adform	52	1.62
teads	51	1.61

Table 6. Selected ad Exchanges. Nodes with out-degree ≥ 50 and in/out degree ratio r in the range $0.7 \leq r \leq 1.7$.

A Appendix

A.1 Selected Ad Exchanges

We select the ad exchanges shown in Table 6 from the *Inclusion* graph by thresholding nodes with out-degree ≥ 50 and in/out degree ratio r in the range $0.7 \leq r \leq 1.7$. One notable ommission from this list is Facebook. The dataset used in this study was collected in December 2015 [10]. Facebook planned the shut down of its public ad exchange around that time [61], which it acquired from LiveRail in 2014 [67].