Network Analysis of Third Party Tracking

User Exposure to Tracking Cookies through Search

Richard Gomer[†], Eduarda Mendes Rodrigues[‡], Natasa Milic-Frayling[*], m.c. Schraefel[†]

[†]Electronics and Computer Science
University of Southampton, UK
{rcg1v07, mc+wi}@soton.ac.uk

[‡]Dept. of Informatics Engineering
University of Porto, Portugal
eduardamr@acm.org

[*]Microsoft Research Ltd
Cambridge UK
natasamf@microsoft.com

***Abstract*—Internet advertisers reach millions of customers through practices that real time tracking of users' online activities. The tracking is conducted by third party ad services engaged by the Web sites to facilitate marketing campaigns. Previous research has investigated tracking practices and tracking agencies associated with popular Web sites. Here we investigate the network properties of the third party referral structures that facilitate gathering of user information for the delivery of personalized ads. By considering third party domains associated with the top ten search results for a diverse set of queries, we arrived at the networks of third party domains in four search markets. We show a consistent network structure across markets, with a dominant connected component that, on average, includes 92.8% of network vertices and 99.8% of the connecting edges. There is 99.5% chance that a user will become tracked by all top 10 trackers within 30 clicks on search results. Finally, the third party networks exhibit properties of the small world networks. This implies a high-level global and local efficiency in spreading the user information and delivering targeted ads.**

Keywords—browser cookies; surveillance; search; network propagation; search queries; trackers

I. INTRODUCTION

Online advertising is the main source of revenue for online businesses [4] and subsidizes most of the free content and services that are available to Web users. Since advertising is most effective when personalized ([2],[19]), the traditional keyword based advertising and contextual banners on Web sites have been replaced by online behavioral targeting that involves real time tracking of individuals' online activities and real time bidding for ad placement on the pages users visit.

This sophisticated advertising approach is enabled by online tracking technologies that are rapidly evolving from cookies to browser fingerprinting and other device identification techniques [3]. As the user visits a site, third party tracking cookies are placed on the user's computer in order to track the user across Web sites. The collected information is used by the ad exchange services that act as brokers between the advertisers who seek opportunities to place advertisements and the Web domain owners, i.e., Web publishers who supply ad spaces on their Web pages (Fig.1). By connecting the two, each ad exchange creates a network comprising Web sites, advertising agencies, and tracking agencies. We refer to these networks as *third party tracking networks*, or *tracking networks* for short.

Previous research has considered the development of tracking practices [7] and the association of tracking domains with popular Web sites [15]. No research to date has considered the networks of trackers that support Online Behavioral advertising (OBA). In this paper we analyze the tracking networks that the user is exposed to when accessing content on the Web through search engines. As the user clicks on a search result, the corresponding Web site may refer to third parties to install tracking cookies on the user machine. By analyzing the referrals of the retrieved sites to the third party domains we derive the tracking networks and their properties. We use a broad set of search topics to analyze search results and tracking practices across different search markets.

Contributions of our research are threefold:

- We provide an in-depth characterization of the tracking networks and show that their structure follows the model of small-world networks where only a small group of entities are highly connected.
- We demonstrate that the referral to third parties from search is independent from search ranking. Essentially, the distribution of trackers across search results is not related to the site's relevance to the user's queries.
- We provide empirical evidence for the extent of user exposure to tracking through search: there is more than 99.5% chance that, in 30 clicks on search results, the user will become tracked by all top 10 most prolific trackers.

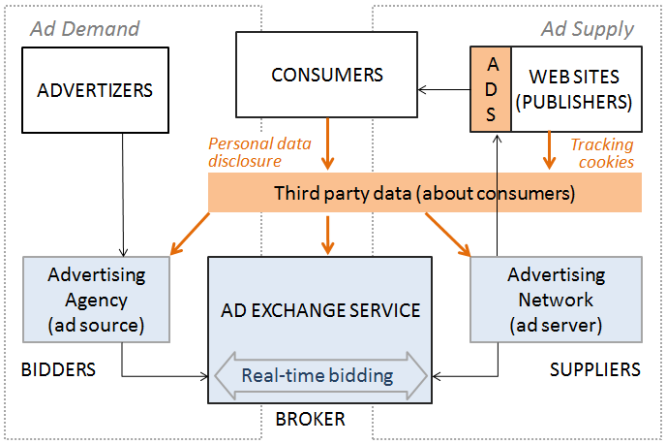


Fig. 1. An adverting ecosystem comprises advertizers who use advertising agencies to bid for ad spaces supplied by Web publishers. Ad exchange services facilitatte bidding on ad spaces. Information about users is shared among the agencies, brokers and ad networks to optimize ad targeting.

In the following sections we first provide the background on the key features of the user tracking practices. We then describe our method of analyzing the tracking networks through monitoring of the HTTP referrer headers. We follow with a detailed analysis of the collected network data and discuss the findings. We conclude with a summary of our contributions and directions for future research.

II. Background and Related Work

A. Online Advertising

Tracking Mechanisms. Web sites who subscribe to an ad exchange service (see Fig. 1), host embedded code (e.g., Java Script) on their pages that connects to the ad networks and loads adverts into the Web page at the time a page is rendered by the browser. In that process, ad networks may store or retrieve cookies containing a persistent user identifier. Such cookies are referred to as *third party cookies*, in contrast to the *first party cookies* that are delivered by the Web site itself. The latter are commonly used to support log-in and multi-page browsing on the site. As the user visits other sites associated with the same ad network, this third party cookie is used by the ad network to identify the user pseudonymously. In this way the ad network obtains, processes, and accumulates data about the user's online activity in real time

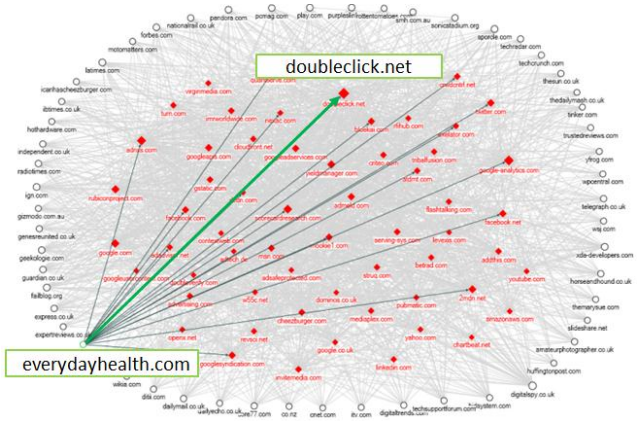
Fig. 2 illustrates the tracking mechanism: shows Web sites a user visited over 10 days (circular nodes) and the third party domains that were referred to during site visits (red diamond nodes). In the user's visit to *everydayhealth.com*, we observe that the visited page referred to a number of third parties which delivered adverts and installed cookies on the user's computer. Among them is the tracking domain *doubleclick.net* which is associated with other Web sites that the user visited (see Fig. 2b). Every time the user visits such a site, the user's action is known to doubleclick.net. That information becomes the basis for behavioral targeting as it captures user's activities across Web sites.

In addition to hosting page elements for the purposes of displaying ads, sites may host "widgets" such as the "*like*" button by facebook.com. These act in a similar manner, allowing a third party, in this instance Facebook, to track Web usage by a particular user across the participating Web sites.

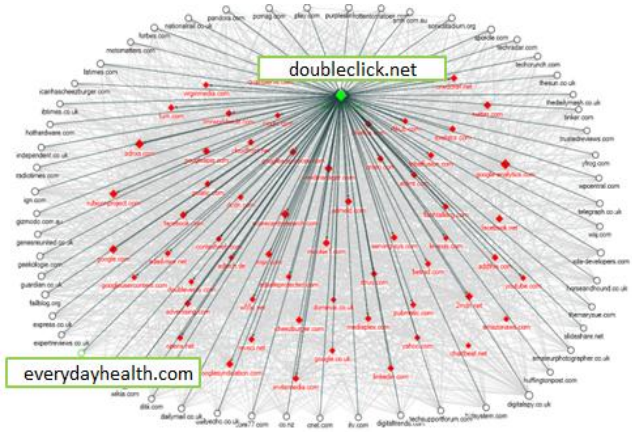
B. Behavioral Tracking and Responses

Online Behavioral Advertising (OBA) aims at inferring users' intent, preferences, habits, and interests from their online activities and selecting personal ads to present to the user. In many instances, the OBA providers, such as audiencescience.com and audiencetargetting.com offer *retargeting of ads* [7]. Ad retargeting involves placing an ad related to the Web site on the pages of subsequently visited sites and extending the exposure to the ad over time.

Privacy concerns related to the user tracking ([10],[20]), led to OBA approaches that reduce the scope of user information that is shared during ad targeting. Among such methods is user modeling on the client side, i.e., within the browser.



(a) Third party domains associated with the *everydayhealth.com*



(b) Network of Web sites and third party domains that are coordinate through *doubleclick.net*

Fig. 2. Network of Web sites (circle nodes) and third party domains (red diamond nodes). A directed edge A→B indicates a 'referral to domain B by domain A'. As the user accesses everydayhealth.com, a number of third party domains are enabled to install cookies on the user's machine. The graph in (b) shows that the doubleclick.net cookie is used across a number of Web sites that the user visits.

The system Adnostic by Toubiana et al. [17] uses a browser extension that incorporates behavioral targeting algorithm based on a local database of browsing history, not shared with external parties. Similar attempts towards privacy protecting techniques have been explored by Langheririch et al. [8] and Tomlin [16]. In 2011, Riderer et al. [14] proposed an alternative mechanism of *transitional privacy*, allowing the user to decide what personal information is released and put on sale while receiving compensation for it.

At the same time, a study of the users' perceptions of OBAs by Ur et al. [18] demonstrated that online users are unaware of the tracking practices and often have incorrect understanding of the role of the advertising networks. Browser add-ons, such as *Ghostery* (www.ghostery.com) and *Collusion* (www.mozilla.org/collusion), aim to reveal the tracking domains and connection of the Web sites to specific trackers. While these are attempts to inform the users about various aspects of the advertising ecosystem there has not been an in-depth qualitative analysis of the tracking networks that result from the ad targeting practices.

C. Analyses of Tracking Practices

Krishnamurthy and Wills [6] examine technical aspects of data aggregation by the third parties and, through longitudinal observations of the techniques and entities involved in the tracking practices, show that the market is consolidating towards strong dominance by a few companies.

Furthermore, Roesner et al. [15] differentiate among 5 third-party tracking practices based on the mechanism they use to manipulate the browser state. They designate them as: (1) *analytics*, for within site monitoring using a third party (e.g., Google Analytics), (2) *vanilla*, for cross site monitoring (e.g., DoubleClick), where a third party stores and aggregates user data, (3) *forced tracking*, for using pop-ups or similar mechanisms to force the users to visit the tracker's site, (4) *referred*, for negotiating with a service or a cross-site tracker to provide the unique user identifier, and (5) *personal*, for embedding a tracker (e.g., Facebook 'Likes' widget) that the user visits directly. Roesner et al. [15] select 1,000 Web sites from the Alexa service (www.alexa.com) to observe the tracking practices and show that on most of the observed sites the users are tracked by multiple parties which combine different tracking practices. The coverage of the Web sites by the trackers varies with few of them playing a dominant role with large coverage.

III. Research Focus and Methods

Our research focusses on the network aspects of online tracking and the exposure of users to third parties through Web services. By considering Web search services we have a realistic scenario in which the user is exposed to a broad range of Web content and a well-defined set of first party domains, i.e., the Web sites in the search results that the user intentionally visits as part of the search task. The latter makes the analysis of the referral and tracking practices more practical and precise. It reduces the ambiguity that may arise from the referral mechanisms. For example, a domain that 'behaves' like a first party but was not included in the search results can be confidently classified as a third party domain.

In our research we aim to answer several key questions:

- How does the search context affect the user's exposure to tracking practices?
- To what extent are the users exposed to the tracking networks through search?
- What are the characteristics of the third party tracking networks?

A. Experiment Design

Search engines are essential for accessing relevant content on the Web. Commercial search services such as Google (google.com), Bing (bing.com), and Baidu (baidu.com), process millions of queries and serve millions of customers each day. However, each access to a search result may lead to a contact with one or more underlying tracking networks. In order to analyze the tracking networks and the user exposure to them through search, we considered search results from Google, Bing, and Baidu and performed a detailed analysis of the referral networks that comprise the retrieved Web sites and the associated tracking domains.

TABLE I. Distribution of search queries across top 12 categories ranked by the number of queries used in the experiments

Category Label	Num of SearchQueries
Shopping\Stores & Products	101
Information\Local & Regional	95
Information\Companies & Industries	60
Living\Health & Fitness	49
Living\Car & Garage	41
Information\Law & Politics	40
Living\Travel & Vacation	39
Living\Fashion & Apparel	37
Information\Science & Technology	36
Living\Finance & Investment	34
Living\Food & Cooking	33
Information\Education	30

B. Search Queries

For search queries we considered a subset of broadly accessible data from the KDD Cup 2005 Challenge [5]. The KDD Cup task involved automated categorization of queries and the organizers published a set of 800 training queries that had been categorized by three human labelers. We selected a subset of the labeled queries, aiming for a consistency in the assignment of queries across the categories and the label accuracy. We excluded labels with no agreement among the labelers and discarded the corresponding queries. The resulting list contains 662 labeled queries. The categories comprise a two level hierarchy with top categories: Computers (8 sub-cat.), Entertainment (9 sub-cat.) Information (8 sub-cat.), Living (18 sub-cat.), Online Community (6 sub-cat.), Shopping (6 sub-cat.), and Sports (11 sub-cat.). Only one second level category had a further sub-category. Each query was assigned between one and four labels from a set of 67 distinct second and third level categories. A list of labels associated with 30 or more queries is shown in Table I. We hypothesize that search results related to the queries in different categories will lead to the retrieval of different types of Web sites which, in turn, may have different marketing practices and involve different third party entities.

C. Search Results

For each of 662 queries we gathered top ten search results from both Google and Bing. We used the Bing search API from the Azure Marketplace to obtain Bing search results while for Google we extracted URLs from the search result pages. The process was carried out four times to collect results for four English-language search markets: United States, United Kingdom, South Africa, and India. Search within a specific market was conducted by providing the

TABLE II. SEARCH MARKETS USED TO COLLECT SEARCH RESULTS

Market	Bing API Identifier	Google Search Domains
India	en-IN	www.google.co.in
South Africa	en-ZA	www.google.co.za
United Kingdom	en-UK	www.google.co.uk
United States	en-US	www.google.com

appropriate location identifier for the Bing API and by altering the search engine domain for Google (see Table II).

Some search queries returned zero results for one or more of the search engine (SE) and market combinations and were excluded from the set for the sake of consistency. That resulted in the final set of 659 search queries, with ten search results for each of the eight SE/market combinations—a total of 5,272 sets of search results comprising 9,776 unique Web domains.

Furthermore, we collected search results from Baidu in the Chinese market by considering popular queries that are published weekly by baidu.com. We compiled a set of 98 queries and collected the top 10 search results for each query. While this data sample is smaller, it enables us to analyze the characteristics of the tracking practices in a non-English language market.

D. Tracking Domains and Cookies

In order to collect information about cookies associated with search results, we adopted the Selenium Web browser automation framework to automate visits to each set of search results. We used the Firefox browser v.16.0 on multiple Linux machines to access Web sites in parallel. Each browser instance was controlled by a Python program. A new browser instance was spawned every 15-30 seconds, depending on the resources available on the several computers that were used. After initial calibration, we applied heuristics to manage the crawls—browser instances that did not complete their processing within 5 minutes were automatically terminated.

We created a new Firefox "profile" for each set of 10 search results to provide a clean environment for depositing cookies. We also installed a custom logging add-on to record the referrer header of all HTTP requests and to log cookie installation, updates, and deletions. No other add-ons or plugins were enabled within the Firefox browser and its "do not track" feature was not enabled. Each URL from a given set of search results was loaded by the browser in the ranked order, starting with the first result. At the end, the browser was directed to a web service to deposit a log file containing cookie and referrer information by means of an HTML form.

E. Analysis Methods and Tools

We conducted two types of analysis: (1) an analysis of the distribution of tracking domains across search results and topic categories and (2) an analysis of the referral network involving Web sites retrieved during search and the third party domains associated with them. To that effect, for each set of search set results (9 sets in total), we analyzed the referrer headers of the browser's HTTP requests as the search result pages were loaded into the Firefox browser. We collected a list of all the Web domains that appear as search results and those that are referred to as third parties. We noticed that some Web domains, such as facebook.com and twitter.com, appear as both the Web sites in the search results and third party domains referred to by other Web sites. That motivated us to differentiate among four types of Web domains:

1) Web sites—Web domains whose pages appear among search results and are not referred to by other sites. They are thought of as the *first party only* domains.

2) Third party only—Web domains that are referred to by Web sites or other third party domains and never appear among search results nor refer to other domains. Such are, for example, googleanalytics.com or ad services that place ads directly on the Web pages.

3) Dual role—Web domains that appear as both first party and third party domains. Example is facebook.com which appears among search results and is referred to by sites that include the Facebook "Likes" widgets.

4) Ad Exchange Service—Web domains that appear only as third parties, i.e., do not appear in search results, and refer to other third party domains. They are *intermediary third parties* that provide a bridge between Web sites and other third parties involved in ad bidding.

From the referral header information we produced an edge-list representing the 'refer to' relationship and created directed network graphs in which A→B corresponds to the fact that the domain B is 'referred by' domain A when the page from A is loaded into the browser. As noted above, in many instances, the domain B is the ad network or ad exchange service that delivers the adverts to the Web page. In other instances, the referral is due to the fact that the Web site is running a script that refers to a monitoring service, collecting statistics of the Web site usage. We loaded the edge and node lists into R (r-project.org) and NodeXL (nodexl.codeplex.com) in order to calculate graph metrics and visualize the referral networks.

IV. ANALYSIS

A. Search Results and Third Parties

For each URL that was visited during the crawls of search results we recorded its search rank among the top 10 search results and the third party domains associated with the URL's domain name. This allows us to analyze the user exposure to the third party tracking as they review the top 10 search results.

1) Trackers and Search Ranks

Fig. 3 shows the average number of third parties associated with search results at a given rank, across search engines and search markets. We note that the number of third parties associated with the search ranks 1 to 3 appears to be slightly lower. We suggest that is due to the prevalence of several sites that are frequently retrieved at those ranks (e.g., wikipedia.org) and are associated with only few third parties. For each of the 20 most frequent third party domains, we calculated the average rank of search results that they are associated with. We collected the statistics for each SE/market combination and

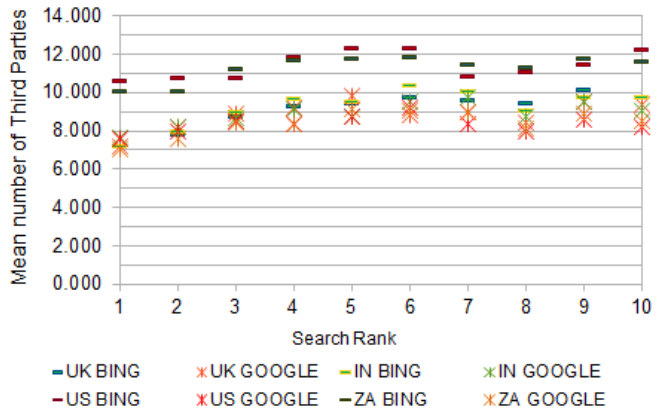


Fig. 3. Distribution of third party cookies across the top ranked search results, in individual search market.

found that none of them differ significantly from the expected average of 5.5. This suggests that the exposure to the top third party domains is equally distributed across all search ranks and the level of tracking encountered by the user is uniform across the search ranks.

2) Distribution of Tracker Types

We sorted the domains of the retrieved Web sites based on the frequency with which they appear among the search result. From each of the 8 search result sets (see Table II) we selected the top 1000 most frequently retrieved Web domains and, after merging the lists, arrived at 3,441 unique Web sites. Among them are 174 Web sites retrieved in all 8 markets. In order to analyze the common tracking practices across the markets we focus on the tracking domains associated with this subset of 174 sites.

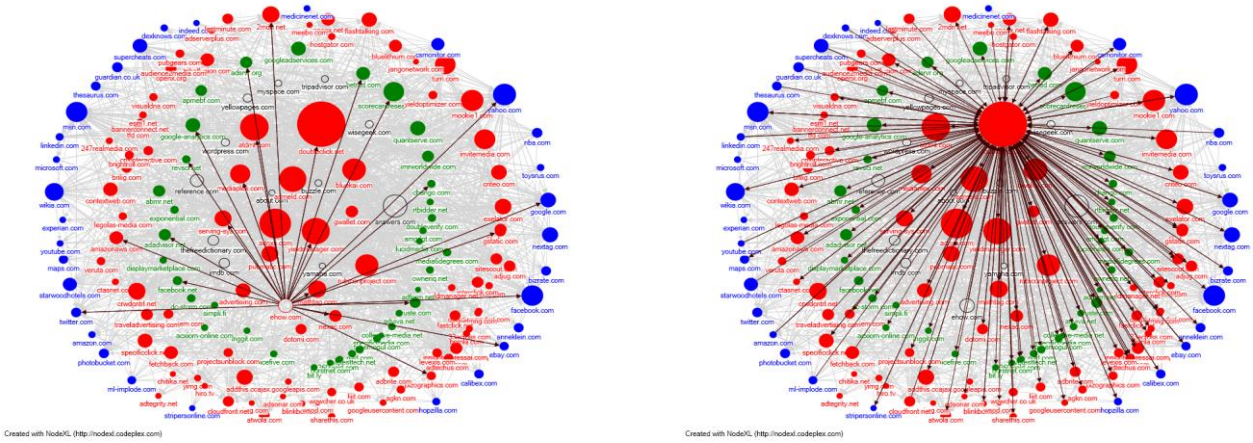
In Fig. 4 we visualize the referral network associated with 30 most retrieved Web sites from the sample of 174. The network includes 714 unique nodes, corresponding to the Web sites and third parties, and 2,311 edges. Further analysis reveals that the network includes is a dominant connected component comprising 711 vertices.

Analysis of the referred third party domains shows that 409 (57%) act as third parties only (see Sec. V). Among the top 10 such domains are: googleadservices.com, google-analytics.com, facebook.net, truste.com, zenfs.com, scorecardresearch.com, quantserve.com, newrelic.com, google.co.uk, and adobe.com. We also identified 58 (8%) of domains with a dual role, i.e., providing an online service and acting as a third party tracker. Among the top 10 are: bbb.org, google.com, facebook.com, yahoo.com, wikipedia.org, twitter.com, indeed.com, linkedin.com, youtube.com, shopstyle.com. We note that the social media sites dominate this list, including the Facebook, Linked In, and Twitter, all of whom have widgets installed by many sites. When the user visits such sites, the trackers are notified and therefore can track the user across all the sites that use such social widgets. The tracking may not be anonymous as sine social media sites, such as facebook.com, know the identity of the person.

Finally, 233 (33%) of 714 domains are ad exchange services and third party tracking companies that enable behavioral targeting. Top 10 companies in this category include: doubleclick.net, googlesyndication.com, gstatic.com, fbcdn.net, tfd.com, invitemedia.com, atdmt.com, ajax.googleapis.com, pinterest.com, and 2mdn.net.

3) Probability of User Tracking

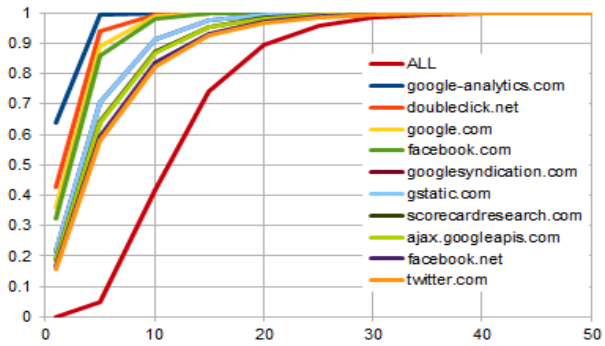
In order to estimate the rate at which users are exposed to third parties, we estimate the probability P(T) that a search result exposes the user to a third-party T by calculating the proportion of search results that refer to T. We rank third parties based on P(T) and, for the top 10, determine the likelihood that the user will encounter each of these parties after accessing a number of retrieved search results. We make two simplifying assumptions. First, we assume that any Web page from a given Web site is exposing the user to the same set of trackers. Second, we expect that the user's choice to visit a search result is independent from the previously seen pages. Based on this model we observe the probabilities that a user would have encountered all top ten third parties. We find that after visiting just 30 search results, the probability of getting cookies from all top 10 third party domains is 99.5%. Fig. 5



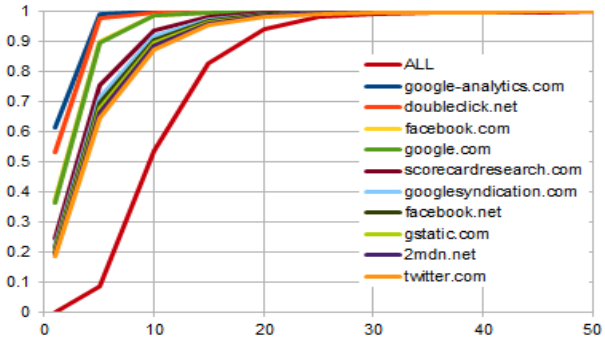
(a) ehow.com site and the tracking entities that are referred to when its pages are rendered in the browser.

(b) doubleclick.net is an ad exchange service, connected with web sites, trackers and ad services.

Fig. 4. Third party network associated with 30 most frequent Web sites among 174 web sites found in all four English search markets, for both Bing and Google. Legend: black circles=Web sites, blue=Dual Role, green=Third party only, red=Ad Exchange/Trackers.



(a) Probability of encountering tracking domains while browsing results of Google in the US Search market



(b) Probability of encountering tracking domains while browsing the results Bing in the US market

Fig. 5. When browsing search results , the user is exposed to tracking domains. We calculate the probability that the user encounters top tracking domains while visiting a random set of search result pages.

shows the probabilities of encountering the trackers when using Bing and Google search engines in the US Market. Our analysis of the Chinese search market indicates that the user exposure rate may be even higher than in the English markets. However, the Chinese dataset is smaller and not directly comparable.

4) Influence of Search Topics

Categories associated with the search queries enable us to observe the users' exposure to tracking while accessing information within a specific area of interest. Because of the limited space, we here present analysis of queries associated with the top ten most commonly applied labels to our query set (see Table I). In particular, we compute the expected exposure to tracking when the user selects the top ranked search result for a given query. Although basic, this model is instructive. As shown in Table III, most of the topic labels led to similar levels of tracking. However, two labels among them show lower numbers of both the third parties that set cookies in the browser and the third parties that do not.

B. Network Analysis

We analyze the tracking network that emerges in each search market and the aggregated global network that comprises all the Web sites and associated third party domains. Table IV presents global graph metrics that highlight the similarity of the tracking networks across the search engines and the search markets.

TABLE III. AVERAGE NUMBER OF THIRD PARTIES ON FIRST SEARCH RESULT BY SEARCH QUERY LABEL. (STD. DEV.)

Label	**Num. of Logs**	**TPs w/ Cookies**	**TPs w/ No Cookies**
Shopping\Stores & Products	785	2.79 (3.57)	3.65 (3.37)
Information\Local & Regional	726	2.16 (4.26)	3.44 (4.27)
Info\Companies & Industries	459	2.88 (4.02)	3.79 (4.28)
Living\Health & Fitness	362	2.33 (3.54)	3.42 (3.42)
Living\Car & Garage	286	3.11 (4.05)	3.84 (3.98)
Information\Law & Politics	298	0.44 (1.26)	1.23 (1.42)
Living\Travel & Vacation	301	3.10 (4.85)	3.19 (2.93)
Living\Fashion & Apparel	289	3.37 (3.87)	3.91 (3.31)
Information\Scienc e & Tech..	271	1.77 (3.16)	2.07 (2.35)
Living\Finance & Investment	245	3.08 (3.68)	3.99 (4.28)

We analyze the tracking network that emerges in each search market and the aggregated global network that comprises all the Web sites and associated third party domains. Table IV presents global graph metrics that highlight the similarity of the tracking networks across the search engines and the search markets.

First, all the networks include a dominant network component that, in some instances, includes more than 92% of nodes and 99% of edges. Second, we compare the tracking networks with synthetic networks based on the Watt-Strogatz random model ([12],[13]). For each network we generate a graph with the same number of nodes and the same average degree of the nodes. For example, a synthetic network for the global tracking network uses 8.91 for the average degree of a node. We then compare the synthetic and real networks based on their average path and the clustering coefficient. As Fig. 6 illustrates, the global tracking network closely follows the *small world network* model with the rewiring probability of 0.2. We get similar results for the tracking networks affiliated with individual search market.

As Latora and Marchiori [9] have shown, small world networks have high local and global efficiency in supporting exchanges of information. Thus, the tracking networks are well equipped to support a range of processes: gathering and disseminating contextual information about the user, real time processing and aggregating information, and bidding and delivery of adverts. We use the same analysis to assess the robustness of the network. Considering the dominant role of the ad exchange domains such as doubleclick.net, we ask how properties of the network would be affected should such a node be excluded from the network. As Fig. 6b shows, removal of a DoubleClick node would increase randomness due to the

TABLE IV. GLOBAL CHARACTERISTICS OF THE TRACKING NETWORKS ACROSS SEARCH MARKETS

	Google				Bing				Baidu	All
Tracking network (G)	**US**	**UK**	**S. Af.**	**IN**	**US**	**UK**	**S. Af.**	**IN**	**CN**	**Global**
Nodes N(G)	5958	6171	5991	6000	5850	6638	5938	6321	473	9733
Edges E(G)	67739	73374	70411	66038	79214	81015	80171	79243	4868	115404
Unique edges E'(G)	26203	26552	25763	26058	25951	28047	26061	26625	1117	43362
Avg. path length L(G)	3.6725	3.3676	3.9216	4.1498	3.2466	3.5300	3.3476	3.2971	3.2918	4.1106
Clustering coeff. C(G)	0.1958	0.1947	0.1993	0.2078	0.2105	0.1818	0.2053	0.2082	0.1685	0.2106
Avg. node degree d(G)	8.7959	8.6054	8.6006	8.6860	8.8721	8.4504	8.7777	8.4243	4.7230	8.9103
Connected components	405	381	398	402	358	436	405	461	12	627
Giant component (GC)	**US**	**UK**	**S. Af.**	**IN**	**US**	**UK**	**S. Af.**	**IN**	**CN**	**Global**
Nodes N(GC) /N(G)	0.9226	0.9316	0.9269	0.9240	0.9313	0.9259	0.9215	0.9155	0.9683	0.9279
Edges E'(GC) /E'(G)	0.9978	0.9984	0.9984	0.9979	0.9982	0.9980	0.9976	0.9972	0.9964	0.9982
Density D(GC)	0.0009	0.0008	0.0008	0.0008	0.0009	0.0007	0.0009	0.0008	0.0053	0.0005

TABLE IV. AVERAGE PATH LENGTH AND CLUSTERING COEFFICIENT OF THE TRACKING NETWORKS ACROSS SEARCH MARKETS, WITH AND WITHOUT DOUBLECLICK.NET

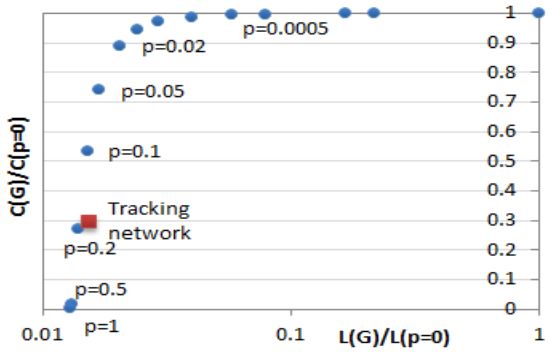
	Google				Bing				Baidu	All
Tracking network (G)	**US**	**UK**	**S. Af.**	**IN**	**US**	**UK**	**S. Af.**	**IN**	**CN**	**Global**
L(G)	3.6725	3.3676	3.9216	4.1498	3.2466	3.5300	3.3476	3.2971	3.2918	**4.1106**
C(G)	0.1958	0.1947	0.1993	0.2078	0.2105	0.1818	0.2053	0.2082	0.1685	**0.2106**
L(G')–doubleclick.net	4.2162	3.9259	4.5185	5.2332	3.8518	4.6238	3.9279	3.8155	3.3148	**5.2152**
C(G')–doubleclick.net	0.1393	0.1407	0.1410	0.1482	0.1492	0.1374	0.1435	0.1440	0.1660	**0.1541**

longer paths and lower clusterability. This, in turn, would reduce the efficacy of the transactions within the network.

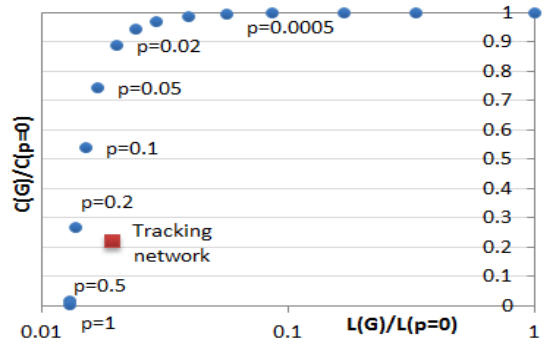
V. DISCUSSION

Our analysis of the search results in different search markets reveals a consistently high extent of third party tracking that supports the OBA. Similarly to the previous studies ([6],[15]), we have identified a small number agencies that dominate the user tracking and advertising markets. However, in contrast to Roesner et al. [15] who focus on popular Web sites and tracking classification based on the mechanisms for implementing cookies, we aim at the user exposure to tracking in a real usage scenario and broaden the scope of the tracking analysis to the referral network properties. Key to our approach is the precise characterization of referrals to arrive at an accurate network representation of the parties involved. One of the challenges is the ambiguity in the roles that Web domains play. Classifying Web domains into first and third parties is too simplistic since the role of the Web domain may change with the context. Fortunately, in the search scenario we can apply the heuristics that a referral to a domain that does not appear among search results can be classified as a referral to a *third party only* domain, e.g., a site analytics service or an ad placement agency.

By considering directional links of the referral networks we can easily observe the role of the domains as illustrated in Fig. 7. An in-link to a node is a referral that often signals to that entity to install an advert and a tracking cookie.



(a) Global tracking network, comprising third party domains from the English search markets. Plotted relative to the Watts-Strogatz random model.



(b) Global network without doubleclick.net.

Fig. 6. Watts-Strogatz random model. The average path length L(G) and the clustering coefficient C(G) of a) the global tracking network; b) the global tracking network after removing doubleclick.net.

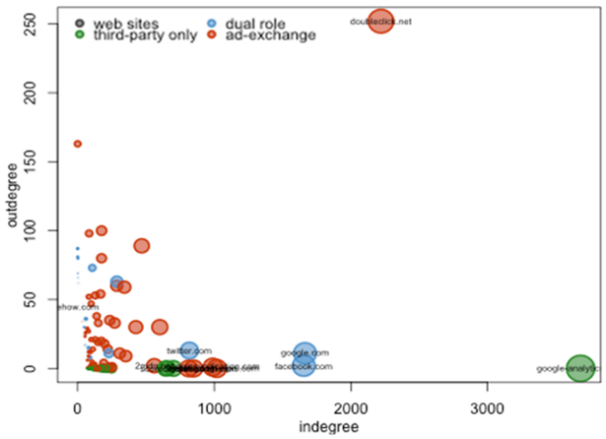


Fig. 7. In-degree vs. out-degree of the nodes of the global network. The size of the nodes is mapped to the total degree.

Most of the out-links are from Web sites to the ad exchange services and ad servers. Domains like doubleclick.net have a large number of in-links and out-links, receiving in-link referrals from the Web sites that are participating in advertisements, and initiating out-link referrals to ad agencies that have been successful in bidding for ad spaces and won the right to place ads on the Web site page. While domains like doubleclick.net are not accessed by end users, the social networks like Facebook and Twitter feature as both the Web site and the tracking domains with similar in-links and out-links characteristics.

The referral networks analysis also reveals the global properties of the tracking practices. We confirmed that the networks associated with the individual search markets as well as the aggregated network across markets follow the model of the small-world networks. Known for their global and local efficiency [9], the small world tracking networks are ideal for distributing gathered data and dispatching the adverts. Thus, the tracking markets are likely to be optimally used. However, due to dominant entities with disproportionally large connectivity, the small world aspect of the network is vulnerable and with that the network efficiencies. At the same time, the prolific presence of a small number of third party entities across Web sites increases the rate of user exposure to tracking. We have shown that a user is likely to be summoned by all top 10 trackers in less than 30 clicks online.

VI. Concluding Remarks

In this paper we perform analysis of referral networks to characterize online tracking practices. The approach enables us to observe the activities of involved parties, i.e., the Web sites hosting the ads, the ad suppliers and the ad exchange services. It provides empirical evidence for the advertising ecosystem depicted in Fig. 1. Since we base our analysis exclusively on the referral detected in the browser, we cannot detect the bidding activities that are conducted within ad exchange services. However, we can observe when the ads and the cookies are installed and by who. The referral networks thus provide an insightful map of the tracking ecosystem. They confirm high efficiency in exchanging information among third parties and capturing users as they browse Web search results. We expect that our approach will be effective in exploring user tracking associated with other types of online services such as content sharing and communication and that derived insights will be useful in informing economic models and policies related to the tracking practices.

References

[1] Amiri A., Menon S.: Efficient scheduling of Internet banner advertisements. ACM Transactions on Internet Technology, Vol. 3, No. 4 (2003) 334–346.

[2] Beales, H. (2011). The value of behavioral targeting. Mimeo, George Washington University.

[3] Eckersley, P. 2009. How unique is your web browser? EFF report, Electronic Frontier Foundation.

[4] Evans, D.S. 2009. The online advertising industry: economics, evolution, and privacy. Journal of Economic Perspectives. April 2009.

[5] KDD Cup'05 Challenge, www.sigkdd.org/kdd2005/kddcup.html#ldata.

[6] Krishnamurthy, B. and Wills, C.. 2009. Privacy diffusion on the web: a longitudinal perspective. In Proceedings of the World Wide Web Conference (WWW'09), 541-550.

[7] Lambrecht, A. and Tucker, C. 2012. When does retargeting work? Timing information specificity in online advertising. MSI Working Paper 11-105. December 21, 2012. Available at SSRN: http://ssrn.com/abstract=1795105 or http://dx.doi.org/10.2139/ssrn.1795105.

[8] Langheinrich M., Nakamura A., Abe N., Kamba T., Koseki Y. 1999. Unintrusive customization techniques for Web advertising. Computer Networks, Vol. 31 (11-16) 1259-1272.

[9] Latora, V., and Marchiori, M. 2001. Efficient behavior of small-world networks. Phys. Rev. Lett. 87, 198701.

[10] J. Mayer and A. Narayanan, "Do not track - universal Web tracking opt out." Available: http://donottrack.us.

[11] Mobasher B., Dai H., Luo T., Sun Y., Zhu J. 2000. Integrating Web usage and content mining for more effective personalization. EC-Web 2000, LNCS 1875, Springer 156-176.

[12] Newman, M. E. J. and Watts, D. J. 1999. Renormalization group analysis of the small-world network model. Physics Let A 263, 341–346.

[13] Newman, M. E. J. and Watts, D. J.1999. Scaling and percolation in the small-world network model. Physical Review E 60, 7332–7342.

[14] Riederer, C., Erramilli, V., Chaintreau, A., Krishnamurthy, B. and Rodriguez, P.. 2011. For sale: your data by you. In Proceeding of the 10th ACM Workshop on Hot Topics in Networks (HotNets-X).

[15] Roesner, F., Kohno, T., and Wetherall, D. 2012 Detecting and defending against third-party tracking on the Web. The 9th USENIX Symp. on Networked Systems Design and Implementation (NSDI 2012).

[16] Tomlin, J. A. 2000. An entropy approach to unintrusive targeted advertising on the WebComputer Networks, Volume 33, Issues 1–6, June 2000, Pages 767–774.

[17] Toubiana, V., Narayanan, A., Boneh, D., Nissenbaum, H., and Barocas, S. Adnostic: Privacy preserving targeted advertising. In NDSS, 2010.

[18] Ur, B., Leon, P. G., Cranor, L. F., Shay,R. and Wang. Y. 2012. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12). ACM, New York, NY, USA, , Article 4.

[19] Yan, J., N. Liu, G.Wang, W. Zhang, Y. Jiang, and Chen, Z. 2009. How much can behavioral targeting help online advertising? In Proceedings of the 18th Int. Conf. on World Wide Web, (WWW '09) pp. 261-270.

[20] 2009. Directive 2009/136/EC of the European Parliament and of the Council.