Tracking Personal Identifiers Across the Web

Marjan Falahrastegar[1], Hamed Haddadi[1], Steve Uhlig[1], Richard Mortier[3]

[1] Queen Mary University of London
[2] University of Cambridge

Abstract. User tracking has become de facto practice of the Web, however, our understanding of the scale and nature of this practice remains rudimentary. In this paper, we explore the connections amongst all parties of the Web, especially focusing on how trackers share user IDs. Using data collected from both browsing histories of 129 users and active experiments, we identify *user-specific IDs* that we suspect are used to track users. We find a significant amount of ID-sharing practices across different organisations providing various service categories. Our observations reveal that ID-sharing happens in a large scale regardless of the user profile size and profile condition such as logged-in and logged-out. We unexpectedly observe a higher number of ID-sharing domains when user is logged-out. We believe that our work reveals the huge gap between what is known about user tracking and what is done by this complex and important ecosystem.

1 Introduction

The rise in the use of personal data and the application of sophisticated algorithms to track and analyse our online browsing behaviour have caused an increase in the number of different tracking services. These services include third-party advertising and analytics services on the Internet and the mobile web [1–3]. User tracking services build a *user profile* by collecting, aggregating, and correlating an individual's browsing behaviour, demographics and interests. While these services are vital for the online economy, there are complex debates over privacy issues that are caused directly or indirectly by such services (e.g., misusing ad tracker cookies to identify individuals [4]).

These services are not only growing steadily in number [2], but are also evolving in terms of mechanisms and technologies. An example of this trend is the emergence of various user tracking mechanisms such as Flash cookies, ETags re-spawning [5] and canvas fingerprinting [6] in a relatively short period of time.

One of the very important phenomena of the Web ecosystem that has been less explored is the practice of sharing user-specific identifiers (IDs). A few works have highlighted the presence of this practice [2, 7]. Moreover, the authors in [6] introduced a method to identify user-specific IDs. Although we are aware of the existence of this phenomenon, our understanding about the extent of this practice and the nature of the parties involved in user-specific ID sharing is rudimentary.

In the rest of this paper, we explore the characteristics of user ID-sharing groups by analysing the organisational and categorical relation amongst the members of ID-sharing groups (§2). We then investigate the effect of user profile on the presence of ID-sharing groups. We show that users are being tracked regardless of their profile size (e.g., amount of their browsing history) and profile condition (logged-in or logged-out)(§3). After discussing the related work (§4), we provide our conclusions (§5).

2 User Tracking

We start our analysis by exploring the connections between domains when they are aimed to track users. User tracking is a practice by which a domain, either being directly visited by a user or indirectly through third-party trackers, assigns a unique identifier to the user, and shares this identifier with other domains. The parties participating in user tracking are able to aggregate the data collected by other parties in order to construct a comprehensive profile of users. In the rest of this section, we first describe our methodology and dataset, and subsequently explore the size and nature of a user ID-sharing group.

2.1 Methodology and Data Collection

We extended the Lightbeam Firefox plug-in to log all headers of HTTP requests and responses. The plug-in additionally records the country where the user is located (our modified version is available in [8]). The recorded data is delivered automatically to our server using an encrypted connection. While there are various Firefox plug-ins to visualize and block third-party trackers, we chose Lightbeam (Figure 1) because of its interactive and easily understandable user interface. We asked our colleagues and friends to install our plug-in and use Firefox as their main browser for the minimum duration of two weeks. In order to preserve users' privacy we did not record any identifiable information such as the IP address, name or contact information. Additionally, we obtained ethics approval from QMUL ethics committee (code QMREC1416a) before performing our user studies. All our data were obtained between 20 February 2015 until 1 April 2015. In total we had 129 participants from 22 countries across the globe. Our participants have visited 4951 unique websites which include 6568 unique third-party trackers. Table 1 lists the number of our participants in each geographical region.

2.2 Nature of ID-Sharing Groups

To explore user tracking via sharing user-specific identifiers, we first need to determine the identifiers that are likely to be used as *user-specific IDs*: a unique identifier stored in a cookie or embedded as a parameter in a URL. For this purpose, we apply the following rules inspired by Acar *et al.* [6] on all items stored in the cookies and the URL parameters.

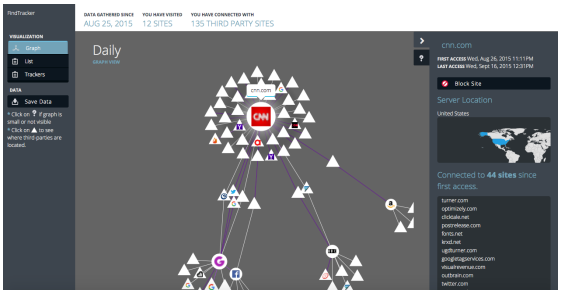


Fig. 1: Lightbeam presents the connection between websites with an interactive and easily understandable user interface.

Region	**Country**	**No. P**
Europe	BE, CH, FE, DE, DK, ES, GB, GR, NL, FI	97
Middle East	BD, IR, QA	11
America	CA, MX, US	9
Oceania-East Asia	AU, CN, MY	8
Africa	SG, MR	3

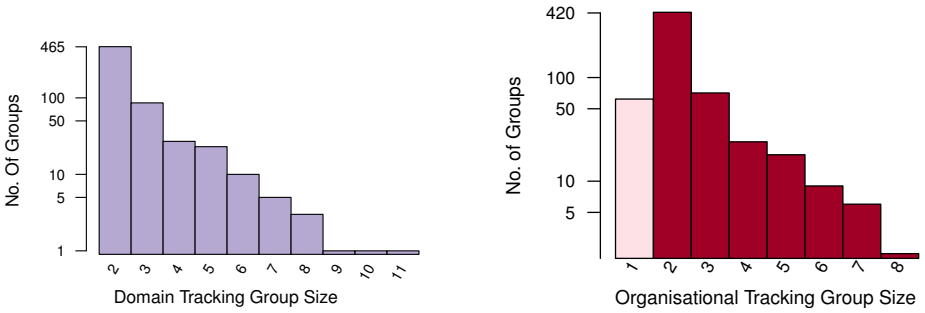
Table 1: Number of participants per geographical location.

- Extract `(key,value)` pairs using delimiters such as ampersand (`&`) and semi-colon (`;`). For instance, this string `id=ece53b2e-ea5c-4433-ad3d&ssid=02ba238451cec44ba88` contains two (key,value) pairs: (id,ece53b2e-ea5c-4433-ad3d) and (ssid,02ba238451cec44ba88).
- Exclude (key,value) pairs that are *inconsistent*: a (key,value) pair is inconsistent if there are multiple values for the same key belonging to a certain domain. For example these pairs (id,ece53b2e-ea5c-4433) and (id,ffc87j3o-gh11-3278) observed from `bbc.co.uk` are excluded.
- Exclude those value strings that are shared by multiple users.
- Only include those value strings that their length is longer than 7 characters. After applying the aforementioned rules on our dataset, we found that 96% of user-specific IDs have a minimum length of 7 characters.

We applied the above-described method for each user. Table 2 shows sample URLs and their identified user-specific IDs with their associated keys. The identified IDs appear in various formats of which the most common are {`xx..x`}, {`x-x-..-x`} and {`x|x|..|x`} where `x` can be any combination of characters and numbers. We find 3,224 unique user IDs from 806 domains. To our surprise, the vast majority of these IDs (96%) are being shared between at least two domains. We identify 769 domains that share unique user IDs with other domains. Extracting the user-specific IDs enables us to identify *user ID-sharing groups*: a set of domains that share user-specific IDs. We identify 660 unique ID-sharing groups

URL	User-Specific IDs	Key
http://ads.rubiconproject.com/ad/11078.js	65d39451-1f73-435a-bf39	put_2760
http://apex.go.sonobi.com/trinity.js	i736hcjtwb05natk	uin_bw
http://cm.adform.net/pixel	d4848|VOzy0|N1xas	adform_pc

Table 2: Example of URLs and the identified user-specific IDs with their associated keys.



(a) Domain ID sharing groups

(b) Organisational ID sharing groups

Fig. 2: Size of ID sharing groups based on number of (a) domains and (b) organisations (Y-axis in both figures uses a logarithmic scale).

containing two to more than eight domains. Figure 2a provides the distribution of the number of different sharing groups (y-axis uses a logarithmic scale) across their group size (x-axis). From Figure 2a, we observe that user IDs are mainly shared between two (467 unique groups, 2742 occurrences) or three (86 unique groups, 201 occurrences) domains. Moreover, the number of unique groups and their occurrences drop steadily as group size increases.

Organisational Sharing User ID-sharing groups consist of multiple domains that may actually belong to the same organisation. Therefore, we broaden our approach from domains to organisations, resulting in *organisational sharing groups*. For example, the organisational sharing group for {`google.com`, `youtube.com`} is {Google}, and for this group: {`youtube.com`, `scorecardresearch.com`} is {Google, comScore}.

To identify the organisation behind a set of domains, we applied a combination of three methods. First, we used Collusion's dataset[3] to detect ID-sharing domains belonging to the same company. We manually inspected this dataset for any changes using websites and wiki pages of the companies involved. Second, we used the e-mail addresses of domains obtained by querying their SOA (Start of Authority) record. The email address, however, is unhelpful if it is a general

[3] `http://collusion.toolness.org/`

account from a cloud, CDN or DNS service. For example, `awsdns-hostmaster@amazon.com` is the email address of all third-parties hosted on Amazon Web Services, and `dns-admin@google.com` is assigned to all services hosted on Google App Engine. We identified the unhelpful email addresses by their email domain name belonging to the known CDN and DNS services, or containing keywords indicating such services. For these cases we used the organization indicated in their `whois` records if available, or else we assumed the domain has no parent company. We are aware that there can be some cases with an outdated `whois` record or email addresses but we believe this is the best approach that can be executed automatically.

Figure 2b provides the distribution of the number of organisational sharing groups (again using a logarithmic y-axis) across their sizes (x-axis). We observe that the number of within-organisational sharing groups (sharing within a single organisation) is considerably lower than those with more than one organisation (sharing across different organisations). Moreover, the most cross-organisational sharing appears between only two organisations. The majority of these two-organisation groups contain a member organisation that appears only once (306). On the other hand, dominant organisations such as Google, Rubicon Project and Optimizely (a user targeting company) appear in 43, 40 and 33 two-organisation groups respectively.

In general, we find some organisations such as Rubicon Project (an ad exchange company) appears strongly in the cross-organisational sharing groups (112 groups) while large organisations such as Google appears in both cross-organisational and within-organisational sharing groups. Table 3 shows the top 15 most popular organisational sharing groups (in their frequency of occurrence) and the nature of their user-specific ID-sharing within the group, i.e., within an organisation (w-org) or cross organisations (c-org).

Cross Categories Sharing To gain more insight into the nature of user ID-sharing, we analysed the ID-sharing groups with a different approach. We examined the categories of domains in each group. We first identified domain categories using the Trend Micro Site Safety Center categorization service[4]. The Trend Micro service contains 85 different interest categories. Moreover, we manually inspected those that were not available on Trend Micro. We find categories related to the ad ecosystem (e.g, ad networks, analytics, ad exchanges) have, expectedly, the highest presence. This strong presence is due to the employed advertising mechanisms (e.g., real-time bidding) that share user-specific IDs across different entities of the ad ecosystem.

We then compared the categories of domains in each group. For instance, in the following ID-sharing group {`getclicky.com`,`ibtimes.co.uk`} the categories of domains in the group are {`Analytics`, `News`}. Table 4 shows the top 15 categories of the sharing groups (in their frequency of occurrence) and the nature of their domain categories in the group, i.e., within a category (w-cat.) or cross categories (c-cat). We observe that the majority of ID-sharing in the

[4] http://global.sitesafety.trendmicro.com

Sharing Group	Type
google.com, googleadservices.com	w-org
google.com, youtube.com	w-org
flickr.com, yahoo.com, yahooapis.com	w-org
bbc.com, effectivemeasure.net	c-org
yahoo.com, yimg.com	w-org
bing.com, live.com	w-org
adxcore.com,cherryssp.net	c-org
rubiconproject.com, wtp101.com	c-org
rubiconproject.com, tapad.com	c-org
bing.com, live.com, msn.com	w-org
eyeviewads.com, rubiconproject.com	c-org
everesttech.net, rubiconproject.com	c-org
rubiconproject.com, w55c.net	c-org
sina.com.cn, weibo.com	w-org
rubiconproject.com, rundsp.com	c-org

Table 3: Top 15 user ID-sharing groups ordered based on their frequency of occurrence. The Type column indicates the nature of organisational sharing within the group (within-organisation=w-org versus cross-organisation=c-org).

groups happens across different categories. We find only 28 ID-sharing groups of which their members belong to the same category (within-category sharing). This number is considerably lower than 110 groups with members belonging to different categories (cross-categories sharing). We have also observed that sensitive domain categories such as health related ones participate in the ID-sharing with domains related to advertisement trackers and search engines (7 groups). For instance, `webmd.com` (a health information website) has shared user-specific IDs with `gravity.com` (an advertisement tracker). Looking at a sample HTTP request from `webmd.com` to `gravity.com` in Table 5, shows that `gravity.com` logs users' visited pages via *referrer* URL-parameter. This information enables `gravity.com` to create users' profiles based on their visited pages and searched terms on `webmd.com`. The presence of such domain categories within sharing groups raises serious privacy concerns since users' sensitive information can be exposed within sharing groups.

3 Effect of User Profile

In the previous section, we observed strong presence of user ID-sharing based on two-weeks online activities' logs of over 100 users. In this section, we further examine the potential intentions behind the ID-sharing by studying the effect of user profile on the presence of ID-sharing domains. For this purpose we run multiple crawls on sets of trained user profiles. In order to create the user profiles, we first created five artificial users with separate accounts on Google, Amazon,

Sharing Group	Type
search engines, web advertisements	c-cat.
search engines, streaming media	c-cat.
ad-tracker	w-cat.
search engines	w-cat.
ad-tracker, web advertisements	c-cat.
ad-tracker, internet infrastructure	c-cat.
ad tracker, photo searches, search engines	c-cat.
media, news	c-cat.
ad tracker, news	c-cat.
web advertisements	w-cat.
ad-tracker, business	c-cat.
health	w-cat.
internet infrastructure, web advertisements	c-cat.
ad tracker, search engines	c-cat.

Table 4: Top 15 categories of the sharing groups ordered based on their frequency of occurrence. The Type column indicates the nature of domain categories within the sharing group (within category=w-cat. versus cross category=c-cat.).

eBay and Twitter. We assigned three different profile sizes, in terms of the browsing histories, to our users: (1) Two users were given a browsing history consisting of Alexa's top 500 websites (*Profile-500*); (2) Two other users with smaller size of browsing history including Alexa's top 200 websites (*Profile-200*); (3) One user with an empty browsing history (*Profile-0*). To explore the effect of not having a user profile, we considered a user with an empty browsing history and without any accounts on the aforementioned websites (*noAcount*). We created the browsing history by crawling the corresponding Alexa's list of websites for five consecutive times while users were logged-in. The profile-training step was done on the Firefox browser installed on a separate Linux machine per user. After creating the user profiles, we installed the Firefox extension from the section 2.1 on the Firefox browsers. Then, we executed the main step of the experiment by visiting Alexa's top 1000 websites for each user. We repeated this step for 20 iterations to expose as many as possible ID-sharing domains. We performed the main step identically under two conditions: user logged-in and user logged-out.

We applied the same rules as described in Section 2.2 to identify user-specific IDs. Consequently, we identified 4,104 unique user-specific IDs shared by 787 domains. Figure 3 illustrates the accumulated number of unique ID-sharing domains across the iterations per user and profile condition. We observe that the highest rise occurs between the first and second iteration (approximately 40%), in comparison with subsequent iterations (Figure 3). Moreover, we explored the number of ID-sharing domains across various profile sizes (browsing histories) and profile conditions (logged-in, logged-out, and noAcount). Table 6 shows the

RequestURL:	`http://rma-api.gravity.com/v1/beacons/log?action=beacon&`*user_guid*`=`**`21737bfabd4416779f6`**`&`*referrer*`=`**`http://www.webmd.com/search/search_results/default.aspx?query=breast-cancer`**
Host:	`rma-api.gravity.com`
Referer:	`http://www.webmd.com/breast-cancer/default.htm`

Table 5: A sample HTTP request from webmd.com (a health information website) to gravity.com (an advertisement tracker). Gravity.com logs users' visited pages via *referrer* URL-parameter. Consequently, the searched terms by users on webmd.com are exposed to gravity.com (e.g. query=breast-cancer)

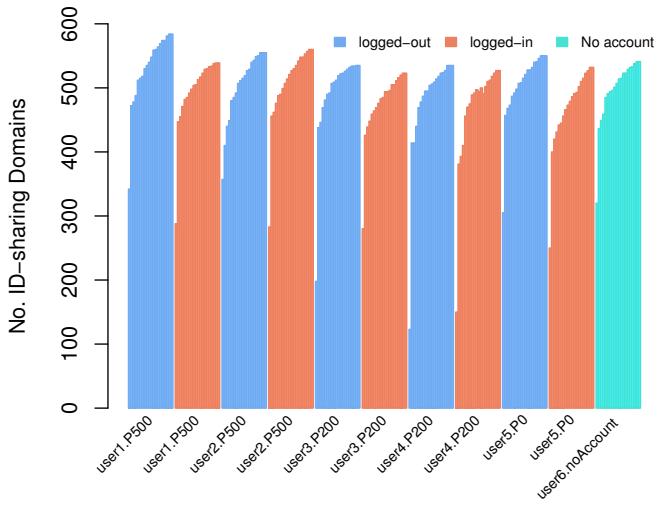


Fig. 3: Number of ID-sharing domains across the iterations. Each bar represents an iteration.

unique number of ID-sharing domains per profile size and condition. The results in Table 6 suggests that users with a larger profile (more browsing history) are tracked by a higher number of ID-sharing domains than those with smaller profile sizes. On the other hand, we find the number of ID-sharing domains, unexpectedly, higher in the logged-out condition than logged-in (Table 6b). In general, the comparable numbers of ID-sharing domains across various profile conditions and profile sizes suggest that the users are being tracked regardless of their profile condition and the amount of browsing history (Table 6).

Afterwards, we examined the presence of organisational ID-sharing groups across different profile conditions. We defined ID-sharing groups as sets of domains that share user-specific IDs (refer to Section 2.2). In addition, we identified the organisations behind the sharing groups using the method described in the Section 2.2. We identified 694 ID-sharing groups of which 357 (=51%)

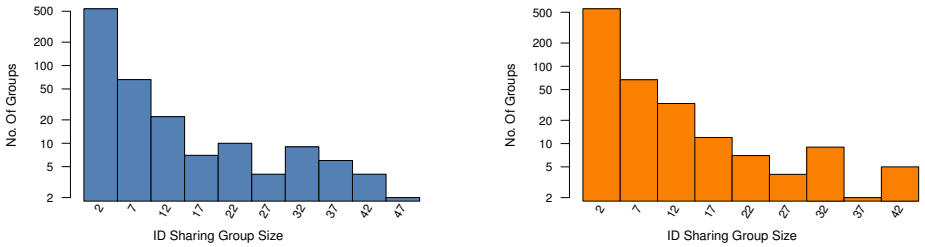
Profile Size	#Domains
P-500	649
P-200	631
P-0	538

(a) Profile Size

Profile Condition	#Domains
no-account	531
logged-in	599
logged-out	749

(b) Profile Condition

Table 6: Total number of unique ID-sharing domains for each (a) profile size and (b) profile condition.



(a) Logged-out (b) Logged-in

Fig. 4: Organisational ID-sharing groups across various profile conditions: (a) logged-out and (b) logged-in (Y-axis in both figures uses a logarithmic scale).

belonging to two distinct organisations. We find that across these groups, Google and Rubicon Project have the highest presence with respectively 27 (=7%), 20 (=5%) cases. Figure 4 shows the number of organisational ID-sharing groups against their group size when the user is logged-out (Figure 4a) and logged-in (Figure 4b). The number of ID-sharing groups with a larger size are higher in the logged-out condition comparing to the logged-in condition. As an example, Figure 5 shows the largest ID-sharing group for the logged-out mode. In this group, we find the Rubicon Project, Switch Concept (an ad. Network company) and StickyADStv (a video publisher company) as the most dominant ones in terms of organisational ID-sharing. We observe strong collaborations between specific organisations such as the Rubicon Project, Sovrn (an ad Network company), Google and StickyADStv.

This unexpected finding can be due to the fact that more domains have been collaborating with each other when the user was logged-out, to compensate for the lack of context about the user, and trying to create a more precise profile for that user—by gathering as much information as possible.

4 Related Work

A number of studies have analyzed trackers from different points of view. Krishnamurthy & Wills [2] showed the expansion of third-party trackers and the

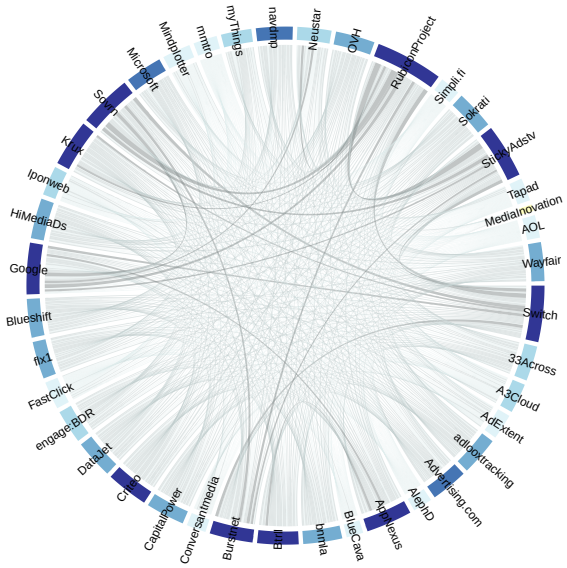


Fig. 5: The biggest organisational ID-sharing group in the logged-out mode. Link thickness represents the frequency of collaboration between two organisations. A Darker colored organisations are involved in higher number of cross-organisational ID-sharing.

acquisitions of tracking companies from 2005 for a period of three years. In [9], they examined the access of web trackers to personal information based on the category of the first-party website in which they are embedded. They found that websites providing health and travel-related services disclose more information to trackers than other types of websites. Gill *et al.* [10] studied the amount of inferred information about users through tracking their visited websites by ad networks. Liu *et al.* [11] have looked at tracking personal data on the web using ISP travel from 2011, however the big shift away from using clear text in the web introduces a much more complicated user ID sharing ecosystem in the web today. They observed that ad networks are able to estimate users' interest with 50% accuracy. These studies showed the possible access of trackers to the user personal information whereas we study the scale and nature of tracking ecosystem.

Roesner *et al.* [12] proposed a framework for classifying the behaviour of web trackers based on the scope of the browsing profile they produce. They show the spread of the identified classes amongst the top 500 websites in the world. Zarras *et al.* [13] studied the ecosystem of ad networks that serve malicious advertisement. Interestingly, they observed some ad networks which more than a third of their traffic belongs to malicious advertisement. Gomer *et al.* [14] focused on the network aspects of third-party trackers which appeared in the

search results of three search engines. They show a consistent network structure of third-party trackers and high efficiency in exchanging information among third-parties.

Mayer *et al.* [15] surveyed different techniques which are used by web trackers to collect user information. Acar *et al.* [6] presented a thorough study of persistent user tracking mechanisms, particularly canvas fingerprinting and evercookies. They introduced a method for identifying persistent user IDs. They crawled top 3,000 Alexa domains, and examined the effect of blocking third-party cookies as well as advertisement opt-out. They observed a decrease in the number of shared IDs, however, they showed that such decrease does not affect the overall access of ID sharing domains to user's browsing history. The main purpose of this study is to explore persistent methods of user tracking through active measurements. Additionally, Olejnik *et al.* [7] studied cookie syncing. They observed the presence of over 100 cookie syncing across top 100 sites. While these studies highlighted the presence of ID-sharing practice across the Web, we focus on the nature of ID sharing groups and their relation with user information using a series of active and passive measurements.

5 Conclusion

In this paper, we explored the entangled connections between all parties of the Web ecosystem. In particular, we investigated the tracking groups that shared user specific identifiers. We recorded the browsing history of more than 100 users for more than two weeks. To our surprise, we find 660 ID-sharing groups in our data. We identify a significant amount of ID-sharing across different organisations. We identified Google and Rubicon Project (an ad. network company) as the most dominant companies that used ID-sharing. Similar to our observation at the organisational level, we observe a significant presence of domains from different categories within ID-sharing groups. We observe that sensitive domain categories such as health related ones participate in the ID-sharing with domains related to advertisement trackers and search engines (seven ID-sharing groups). Moreover, we examined the effect of user profile on the presence of ID-sharing domains. Interestingly, we observe that users are being tracked regardless of their profile condition (logged-in or logged-out) and the amount of browsing history. We unexpectedly observe that the number of ID-sharing domains are higher in the logged-out condition than logged-in. Our results suggest that more domains are collaborating with each other when the user is logged-out trying to create a more precise profile for that user. As a further work, we would like to examine whether this collaboration amongst ID-sharing domains in the logged-out mode aims to identify the user, or it is a side-effect of knowing less about the user, hence being more inclusive in potential advertising sources. Note that from our data we cannot directly observe whether domains use these IDs to merge collected data from different sources. However, considering the possibility of such practice, we believe it is important to get additional insight about what ID-sharing groups actually do through the user IDs.

References

1. N. Vallina-Rodriguez, J. Shah, A. Finamore, Y. Grunenberger, K. Papagiannaki, H. Haddadi, and J. Crowcroft. Breaking for commercials: characterizing mobile advertising. In *Proceedings of the ACM Internet Measurement Conference*, 2012.
2. B. Krishnamurthy and C. Wills. Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th international conference on World Wide Web*. ACM, 2009.
3. M. Falahrastegar, H. Haddadi, S. Uhlig, and R. Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In *Traffic Monitoring and Analysis*. Springer Berlin Heidelberg, 2014.
4. NSA using Google's online ad tracking tools to spy on web users. `http://threatpost.com/nsa-using-google-non-advertising-cookie-to-spy/`.
5. Mika Ayenson, Dietrich J. Wambach, Ashkan Soltani, Nathan Good, and Chris J. Hoofnagle. Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning. *Social Science Research Network Working Paper Series*, 2011.
6. Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 674–689, New York, NY, USA, 2014. ACM.
7. Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pages 199–208, New York, NY, USA, 2011. ACM.
8. Findtracker. `http://www.eecs.qmul.ac.uk/~marjan/repo/findtracker.zip`.
9. B. Krishnamurthy, K. Naryshkin, and C. Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Proceedigs of the Web 2.0 Security and Privacy Workshop*, 2011.
10. Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 141–148, New York, NY, USA, 2013. ACM.
11. Yabing Liu, Han Hee Song, Ignacio Bermudez, Alan Mislove, Mario Baldi, and Alok Tongaonkar. Identifying personal information in internet traffic. In *Proceedings of the 3rd ACM Conference on Online Social Networks (COSN'15)*, Palo Alto, CA, November 2015.
12. F. Roesner, T. Kohno, and D Wetherall. Detecting and defending against third-party tracking on the web. In *USENIX Symposium on Networking Systems Design and Implementation*, 2012.
13. Z. Apostolis, K. Alexandros, S. Gianluca, H. Thorsten, K. Christopher, and V. Giovanni. The dark alleys of madison avenue: Understanding malicious advertisements. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 373–380, New York, NY, USA, 2014. ACM.
14. R. Gomer, E. Rodrigues, N. M. Frayling, and M.C. Schraefel. Network analysis of third party tracking: User exposure to tracking cookies through search. *Web Intelligence and Intelligent Agent Technology*, 1, 2013.
15. Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, 2012.