Big Friend is Watching You: Analyzing Online Social Networks Tracking Capabilities

Abdelberi Chaabane
INRIA, France
chaabane@inria.fr

Mohamed Ali Kaafar
INRIA, France
kaafar@inria.fr

Roksana Boreli
National ICT Australia
University of New South Wales, Australia
roksana.boreli@nicta.com.au

ABSTRACT

In this paper, we examine web user tracking capabilities of the three major global Online Social Networks (OSNs). We study the mechanisms which enable these services to persistently and accurately follow users web activity, and evaluate to which extent this phenomena is spread across the web. Through a study of the top 10K websites, our findings indicate that OSN tracking is diffused among almost all website categories, independently from the content and from the audience. We also evaluate the tracking capabilities in practice and demonstrate by analyzing a real traffic traces that OSNs can reconstruct a significant portion of users web profile and browsing history. We finally provide insights into the relation between the browsing history characteristics and the OSN tracking potential, highlighting the high risk properties.

Categories and Subject Descriptors
D.4.1 [COMPUTERS AND SOCIETY]: Public Policy Issues—Privacy

Keywords
Online Social Networks, Privacy, Tracking

1. INTRODUCTION

The popularity and widespread of Online Social Networks (OSN) has increased in recent years to a level unparalleled by any other Internet service. While these services are freely accessible, the viability of their business model relies on collecting users’ personal data for targeted advertising and in some cases data monetisation [2]. OSN user’s profiles represent a rich source of personal information about the users, including the demographic information, their interests and social relations. The threats to privacy resulting from this direct exposure of personal information have been widely publicized and researched [6]. A second source of privacy diffusion and exposure of information about individuals is the third party tracking of user’s Internet activity i.e. the visited web sites. A number of research studies (e.g. [8]) investigate the mechanisms used for web tracking by third party services including the Ads networks, analytics companies and content distribution networks, and the increasing widespread of such mechanisms on the Internet.

In this paper, we focus on the emerging players in the web tracking arena, the OSN services. The OSN “share” feature provides both an incentive for web site owners to include it on their sites, as it may result in their increased popularity, and a way for OSNs to track user activity on these sites. We investigate in detail the mechanisms for tracking user’s visited web sites and the magnitude (widespread) of the tracking activities for the three most popular OSNs: Facebook (FB), Twitter (TW) and Google+ (G+). We demonstrate how the tracking is done even when the users are logged out of a specific OSN or, surprisingly, when they do not have an OSN account. We further demonstrate the risks from such tracking by an experimental study showing how the majority of user’s web profile can be constructed based on the information tracked by OSNs.

Our contributions include: (i) Highlighting of the mechanisms by which the three most popular OSNs track visited sites and showing that even without an OSN account a user is still vulnerable to tracking. (ii) Evaluation of the widespread of user tracking by the three major OSNs, based on the Alexa top 10000 sites. (iii) Based on real traffic traces, demonstration of the real potential that an OSN provider may construct a large part of user’s web profile; we present examples where up to 77% of user’s web profile is constructed by OSNs only by tracking user’s activity.

We briefly summarize some of the previous research results in the two privacy topics related to our work: information leakage and user tracking. The former may be due to miss-configuration or user miss-understanding of privacy settings, as per [10]. The leakage may also result from the mechanism used by the service to transmit and store private information, where “Personally identifiable information” (PII) may be leaked by OSN providers [9]. Finally, as shown [7], web sites can directly leak visiting user’s private information.

Despite the fact that tracking user activities represents a major threat to user privacy, still it is widely used. The most comparable work to ours is [5] where the authors have studied web activity tracking by third parties. Our work focuses on web activity tracking by OSNs, which was not considered in [8], and presents a comprehensive analysis of tracking by the major OSN service providers.

The authors in [11] have described the mechanism deployed by Facebook to track user navigation. The paper focuses on the legal issues of tracking and provides a succinct technical description of the mechanism. Our work extends their research results by including an in-depth description of how the tracking mechanism is deployed by the three major OSNs, analysing the presence of OSN...
tracking in Alexa top 10000 web sites and providing tracking statistics based on real traffic measurements.

The organization of the paper is as follows. Section 2 presents the techniques deployed by the three majors OSNs to track users. Section 3 describe the widespread of such mechanism in top10K most visited web site as classified by Alexa. Section 4 presents an analysis on real web traffic and demonstrates the efficiency of tracking system on building accurate user profile. We conclude in section 5 and present ideas for future work.

2. PRIVACY ELEMENTS

This section describes the implementation of the OSN Tracking Mechanism (we will call hereafter OTM or bugs interchangeably) by presenting its main components. We then describe each technique of the studied OSNs and evaluate each of their tracking capabilities.

2.1 Preliminary Operations

As illustrated in Figure 1 when a user visits a web site, the fetched page typically generates several other HTTP connections. These connections, which aim to get additional components into the loaded web page, can be classified in two groups: i) connections belonging to the user’s visited web site (usually referred to as first party connections), and used to fetch “legitimate” resources such as images or css (e.g. step 2 in Figure 1). ii) other connections established to third party entities for the purpose of content serving, Ads, tracking, analytics measures and statistics computation or a combination of thereof (step3). As such, these third parties can be Ads Networks (e.g. Google’s Adsense, Yahoo! yieldmanager, AOL) or analytics companies (e.g. Google analytics, Microsoft/aquantive) or content distribution networks (e.g. akamai, yimg), but also, as this papers concentrates on, Online Social Networks (e.g. Facebook, Twitter and Google+).

In the case of an OTM, these connection triggers are embedded into the fetching of the “Share” button resource. In fact, an increasing number of websites provide a button to share their content (e.g., from user’s profile). Cookies are the de facto technique that is typically used to report user’s activities while ensuring unique tracking identity. In this section, we examine specific cookies-based tracking mechanism that we observed for three of the major OSN actors: Facebook, Twitter and Google+

2.2 Transmitting the url of the webpage

When loading the OTM resource the user browser typically transfers the current web-page address to the OSN third party either via the referrer attribute of the HTTP request or as a simple url parameter. For example, when fetching www.cnn.com an HTTP request is sent to Facebook. As shown in the request (below), the url parameter origin specifies the origin of the request transmitted to the OSN (the article that the user is accessing).


Similarly, the referrer record as shown below reports the url of the website the user accesses.

GET www.facebook.com/plugins/likebox[...]

2.3 Cookies-based tracking

In addition to the webpage url, OSNs need to uniquely identify the user to bind this information to potential previously gathered data (e.g., from user’s profile). Cookies are the de facto technique that is typically used to report user’s activities while ensuring unique tracking identity. In this section, we examine specific cookies-based tracking mechanism that we observed for three of the major OSN actors: Facebook, Twitter and Google+

2.3.1 Facebook

Facebook stores 16 different cookies for an active session. The role of each one is hard to determine but two of them are made to identify users: d_atr and c_user. The d_atr cookie contains a random string (24 characters) and is set when a user visits Facebook.com. If a user does not have a Facebook account but has once visited Facebook website, this user will also have this cookie set. The cookie is valid for two years, which suggests that Facebook has the ability to track all users that have once visited Facebook (and even without registering to the service) for a relatively long period of time. This information of the “past” could then theoretically be linked to the user upon registration of an account using a unique web browser.

As shown in Table[1] when a user is logged in, the user’s browser sends both c_user and d_atr to Facebook and hence the user is uniquely identified. When the user logs out, the d_atr cookie is not deleted and the browser keeps on sending it to Facebook when loading OTM resources. By matching the d_atr with c_user, Facebook is able to track user’s activities who have logged out.

To summarize Table[1], entries for Facebook, we note that we observed that Facebook has the ability to track all users that have visited facebook.com at least once, and logging out from the service does not help users to avoid being tracked.

2.3.2 Twitter

Twitter stores 15 different cookies for a logged-in user. Again, although it is difficult to determine the role of each cookie, we note two cookies that are used as a user identifier: twid and guest_id. guest_id is set independently from whether the user is logged-in or not, as opposed to the twid cookie, which is only present during a user’s active session. These two cookies can be used to link the activity of logged-out users with his real Twitter ID. Moreover, Twitter shows an interesting behavior when a user who has never
visited the Twitter website visits a website which contains a Twitter “Share” Button: Twitter automatically set the guest_id cookie and hence may track users that never visited twitter.com. This behavior is unique, since none of the studied OSNs behaves similarly.

2.3.3 Google+

Google stores 28 different cookies for a logged-in user. These cookies belong to multiple domains and sub-domains of Google. Multiple cookies are used to uniquely identify user, among which a few are sent through an encrypted connections (e.g. SSID). We identified and studied two of those that are sent in clear, SID and PREF. The latter contains a random string variable ID that uniquely identifies the user.

When the user disconnects, the PREF cookie is not deleted and is still transmitted to Google when the browser loads a Google+ OTM resource. As such, theoretically, Google is still able to track the web activities of logged-out users. To achieve this, Google has to match the ID value from the PREF cookie to the user’s profile, which is possible since both the session-id carrying SID cookie and the ID carrying PREF cookie are together visible to Google while the user is logged-in.

Summary

An interesting finding when digging into the cookies-based OTM is that all the studied OSNs provide a mechanism (theoretically capable of) tracking users who have explicitly logged-out from the service. Moreover, users that did not register into these OSN services are still trackable, since OTM allows also to keep web browsing traces of users prior to their registration to the system, assuming they visited the OSN’s domains once. Twitter exhibits a unique, yet surprising tracking behaviour and is theoretically designed to track users that do not have accounts, and never visited the twitter domain either.

3. OTM COVERAGE OF ALEXA TOP 10000 SITES

In this section, we analyze to which extent OTM is diffused across the web. We consider the top 10K most popular websites in which we evaluate the OTM distribution. We also compare OSNs tracking possibilities to Google tracking mechanism.

We extracted the top 10200 most visited websites according to Alexa[1] from which we removed 128 entries belonging to search engines. The remaining list was crawled and all subsequent HTTP requests recorded. We further removed 138 web pages containing Javascript and/or flash programming bugs. The final list of analyzed websites contains a total of 9933 different urls (which we will refer hereafter as top10k websites), out of which 7173 representing 72% of the considered websites include at least one tracking element in their main page.

We used a modified version of Ghostery[2] to tag tracking and ad companies.

Google’s tracking system is embedded in nearly half of the websites (43.29%). This presence is mainly due to Google Analytics and Google AdSense. Facebook with both its “Like” button and “Facebook Connect” plugins is second with a surprising coverage of 22.2%. Google+ and Twitter are represented by 10.4% and 7.25% of the webpages, respectively.

3.1 Ranking Coverage

As a first step to measure the diffusion of OTM, we draw the distribution of the number of detected OTMs as function of the website rank in Figure 2. This distribution suggests a uniform diffusion of OTMs through the top10k websites. As observed previously, Facebook has the highest coverage. In particular, it covers nearly 500 websites out of the top 2000 websites whereas Twitter and Google+ cover only 145 and 168 respectively.

Surprisingly, these OTMs are not concentrated in most visited websites. In particular, among the top 1000 sites, only a few dozen of sites embed OTMs. Two reasons may explain this observation. First, the sites geographical origin may have an impact on such a distribution, since the studied OSNs are not as spread in Asia as in Europe and U.S (see Figure 3). Second, the website content and categories may be not suitable to embed such OTMs, since OTM main targets are more likely to be interactive online services such as Blogs or News. From a geographic perspective, OTMs target all origins as shown in Figure 3. In this Figure, the y-axis represents the percentage of websites belonging to the continent (illustrated in the x-axis) that contain at least an OTM. For example, 50% of

---


[2] We used en.wikipedia.org/wiki/List_of_acquisitions_by_Google to tag all services that belong to Google

[3] We define the site origin as the origin of its main audience (i.e. The geographic location of the majority of its users)
In this section, we study the distribution of OTMs according to the category of the website. We used [4] to extract the websites categories (e.g., cnn.com is categorized as News). We retrieved 81 categories from the top 10K websites, and then computed the distribution of OTMs as a function of these categories.

We observe that among the total set of categories, Facebook covers the majority with 68 categories (83%), the most prominent being General News (14%), Entertainment (11.6), Internet services (5.8%) and Online shopping (5.5%). Similarly, Twitter covers 60 different categories, where the most illustrated categories are Entertainment (12.5%), General News (12.2%), Internet Services (7%) and Blogs (5.6%). Finally, Google+ is also covering almost 80% of the categories (64), but interestingly it is well represented in different categories such as Pornography (5.6%).

This shows that OTMs are widely spread and not restricted to a small number of categories. This confirms that if collected, OTM-based information would depict an accurate user profile. Hence, the collected data rise concerns about users privacy, since it is collected without neither the user consent nor his knowledge.

3.2 Category Distribution

In this section, we study the diffusion of OTMs according to the continent. We used [4] to extract the continent-based origin. North American websites in our dataset contain a Google tracking mechanism whereas this coverage reaches only 22% of Asian websites. Not surprisingly, Asia is poorly targeted by the studied OSNs, which can be explained by the poor popularity of these in Asia due to the existence of local concurrent services (Cyworld in south Korea, or Sina Microblog in China, etc.).

The general trend observed above is still valid, with Google tracking system being the most represented through all the continents, and Facebook leading the OTM diffusion independently of the continent.

3.3 SSL-based connections

Transmitting cookies to OSN servers may raise security concerns if the connection is insecure (e.g., unencrypted WiFi connection). Hence, transmitting these cookies over SSL is preferable. We observed that Google+ uses SSL in the vast majority of the cases, and only a few requests (16%) are still sent in clear. Twitter and Facebook OTMs differ significantly from Google+’s behaviour with 96% and 96.5% of the traffic containing the transmitted cookies sent in clear, respectively. This behaviour may endanger the privacy and security of users since an attacker may easily perform a session hijacking.

4. ANALYSING REAL TRAFFIC TRACES

So far, we have studied the diffusion of OTMs, and observed that they are embedded in a wide range of websites and hence can be effectively used to track users outside the OSN sphere. Now, we concentrate on quantifying how much information is actually collected in practice.

4.1 Used Dataset

We captured all the HTTP headers (requests and responses) transiting through our lab local network, for a period of a week starting from the 20th of October 2011. We anonymized the observed traffic by hashing the source IP, and constructed two datasets.

- **Dataset1** containing all the traffic with 687 different IP addresses. It contains more than 27 million connections to nearly 55K different destinations.
- **Dataset2** For the purpose of our experiments in Section 4.3, we reduced Dataset1 so that the number of overall destinations is reasonably low. In fact, since we are classifying websites into categories, we were unable to categorize all the 55K destinations and hence chose to randomly sample a subset of users from Dataset1. Reducing the number of users allows us to keep the history characteristics (length, diversity) whereas reducing the number of destinations to classify. This reduced dataset, called Dataset2 contains 69 IP addresses that made over 16 million connections to 17539 different destination servers. We further filtered out Ads servers and static content providers, which reduced the set of destinations to 5712 different addresses.

4.2 Who’s connected?

In the following, we use Dataset1 to estimate the proportion of users logged-in to the considered OSN services. We capture the transmitted cookies to estimate such statistics. Although, as mentioned in section 2, cookies are not reliable to determine whether a user has an account or not, establishing whether a user is connected to the service by observing the session cookies still fit for purpose, since the later is transmitted by the browser only when the user is logged-in.

We observed that 48% of the users in our dataset have a Facebook account and 33% were logged-in at least once during our measurements. The number of users with a Twitter account is significantly lower than the number of Facebook users, with only 195 users (28% of the total number of unique IP addresses) observed to have an account an only a small fraction (4%) being logged-in at least once.

We do not present results for Google+ since we did not observe a significant number of cookies transmitted in clear, due to the usage of SSL by default in most of Google services.

The relatively small proportion of logged-in users we observed, may be explained by the IT awareness of users we monitored in our dataset. Most users here are computer scientists and thus aware of potential tracking threats, and may use different techniques to mitigate such threat by using cookies blocking/cleaning plugins, enforcing HTTPS connections or using private mode browsing.

4.3 OSN profiling

So far, we showed that theoretically, OTMs are an efficient way of tracking users. In this section, we aim to study to which extent
Figure 4: From left to right: History size VS History coverage(%), History size VS History coverage(#), History size VS Profile coverage

this mechanism can be used to “construct” users’ profiles. From dataset2, we construct simple profiles based on either the web history of the user or on the visited websites’ categories and then observe how much of these profiles can be reconstructed by the OSN.

4.3.1 Methodology

Most profiling techniques rely on classifying websites into categories to overcome the huge amount of information generated by user’s web browsing. Based on these categories, a user Profile can be constructed [3].

For the sake of profile construction simplicity, we define a user profile $P_u$ as the union of all websites categories visited by a user $u$. From dataset2, we extract all destination hosts and used [4] to categorize the visited websites, 98.44% of which have been successfully classified. Note that in our experimentation a website may be classified into up to three different categories.

Each website visited by a user $u$ belongs to at least one category $c_j (1 \leq j \leq C)$ (the total number of observed categories is 81 in our dataset). The set of all visited categories represent the user profile ($P_u$) whereas the set of visited websites (i.e., the user web history) is denoted by $H_u$. Furthermore, $HC_u$ denotes the proportion of $H_u$ retrieved by either Facebook or Twitter using their respective OTMs. $PC_u$ represents the fraction of $P_u$ that Facebook (resp. Twitter) is collecting through the OTM. For instance, if a user visits cnn.com, kernel.org and foxnews.com, cnn.com and foxnews.com are classified as News and Kernel.org as Software. Since only cnn.com has a Facebook OTM, $PC_u$ is 50% whereas $HC_u$ is 33.3%. Figure 5b shows the distribution of users according to the number of categories in their profiles. We observe that 75% of the profiles have between 10 and 40 different categories.

Figure 5b shows the CDF of $HC_u$. For example, more than 50% of users have at least 15% of their web history contain an OTM from both Facebook and Twitter. It also shows that at most 10% have more than 25% of their web history tracked by the two OSNs. A surprising result is that the coverage of Twitter and Facebook are slightly different as opposed to our results observed in Section 3.1 (Alexa case).

Second, we analyzed the relation that might exist between the size of web history $|H_u|$ and $HC_u$. Figures 4c and 4d depict our finding. Most users have $HC_u$ between 10% and 25% whereas we observe a large variation of $HC_u$ for small history size. Thus, a small history size does not necessarily entail a small coverage but rather depends on the category of visited websites. Moreover, as $|H_u|$ increases, so does $HC_u$. Thereby, users with large history sizes tend to be more “efficiently” tracked than others.

Finally, we analysed the correlation between the history size $|H_u|$ and the profile coverage $PC_u$. As shown by Figure 4b, the larger $|H_u|$ is, the higher $PC_u$ is. These two last results were expected since large history size implies that some of websites contain an OTM with high probability. A second observation is the large variation of $PC_u$ for users having small history sizes. For instance, different users with $|H_u| = 40$, can have as different $PC_u$ value as 0, 22%, 30% or 68%. As for the $HC_u$ case, a small history size does not necessarily implies a small profile coverage. On the other hand, this variation decreases with a larger history size. In fact, most users with $|H_u|$ larger than 200 have a $PC_u$ value higher than 40%.

4.3.3 User Profile analysis

Previously, we showed that users with larger web history tend to be more tracked than others. Nonetheless, we also note that even users having small history size are still tracked but with a larger variation. In this section, we examine the correlation that might exist between the profile size $|P_u|$ and the profile coverage $PC_u$, which is depicted in Figure 4d. It indicates a similar trend to the relation observed for the history size. In essence, users with a large $|P_u|$ tend to be more easily trackable whereas other users with a small $|P_u|$ exhibit a larger variation of $PC_u$. However, for the vast majority of users, Facebook and Twitter are able to reconstruct a very accurate category-based profile. In fact, $PC_u$ varies from 0.4 to 0.77.

To dig more into the reasons that would explain our findings, we ask whether users with a high $PC_u$ value browse similar web content (i.e., similar categories). To answer this question, we compute a similarity matrix as follows: for all users, we extract the set of all $P_u$ such that $PC_u \geq \beta$, where $\beta$ defines the level of reconstructed profile coverage. Then, we compute the Jaccard index between all profiles in that set. Recall that the Jaccard index is com-
between users who are highly tracked. For instance, all users who and green (##\text{...}##) are considered highly sensitive as it may refer to users health, is being reported to Facebook. Among the websites in our Alexa dataset, 34.4% of the websites classified as potentially providing Health information do embed at least one OTM.

This paper presents insights about a new tracking mechanism that can be used by OSN providers to collect web browsing information about their current users, and, as we have demonstrated, about potential future users. We observed that these increasingly popular mechanisms cover a broad range of content ranging from Blogs, Health to Government websites. Specifically, we draw attention to the potential privacy threat that may arise from this accumulation of private data that may be utilised for user profiling, without the user consent. We showed that this data, if collected, can draw an accurate profile of the user interests.

7. REFERENCES


5. DISCUSSION

Since OSNs do already maintain a tremendous amount of information about their users, it may be argued that tracking mechanisms could barely affect users privacy. Nonetheless, many websites when visited may leak sensitive and personal information about users. For instance, when a user visits [www.sparkpeople.com](http://www.sparkpeople.com) he might not be aware that such an information about his web browsing activity, which can be considered highly sensitive as it may refer to users health, is being reported to Facebook. Among the websites in our Alexa dataset, 34.4% of the websites classified as potentially providing Health information do embed at least one OTM.

On the other hand, users’ navigation patterns can be used to divide customers with similar preferences into several segments through web usage mining and then recommended targeted ads as shown in [5].

Since OTM are cookie-based tracking mechanism, solutions such as cookies deletion after each session or private mode browsing can be a first approach to mitigate the threat. However these countermeasure suffer from at least two drawbacks. First, navigation patterns belonging to one session are still tracked. Second, more aggressive tracking techniques (e.g. through IP address and browser fingerprint [12]) can still be applied. A more suitable solution consists in blocking the OSN “iframes” connections, and as such all connections to the OSN servers. Tools like Ghostry [1] implement such a mechanism as a browser Plugin.

6. CONCLUSION

Figure 6: $P_u$ similarity using Jaccard index (FB left, TW right)