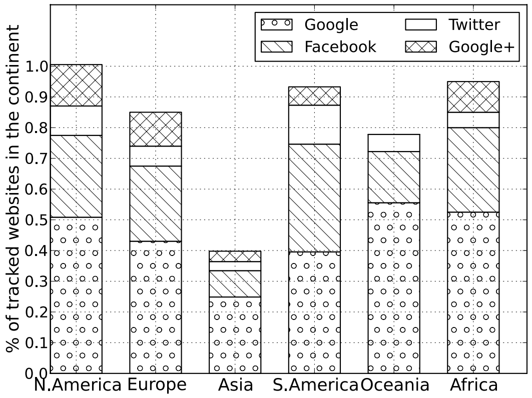

Figure 3: Proportions of tracking mechanisms per webpages continent-based origin.

ing 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.

3.2 Category Distribution

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.

3.3 SSL-based connections

Transmitting cookies to OSN servers may rise 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 my 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 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.

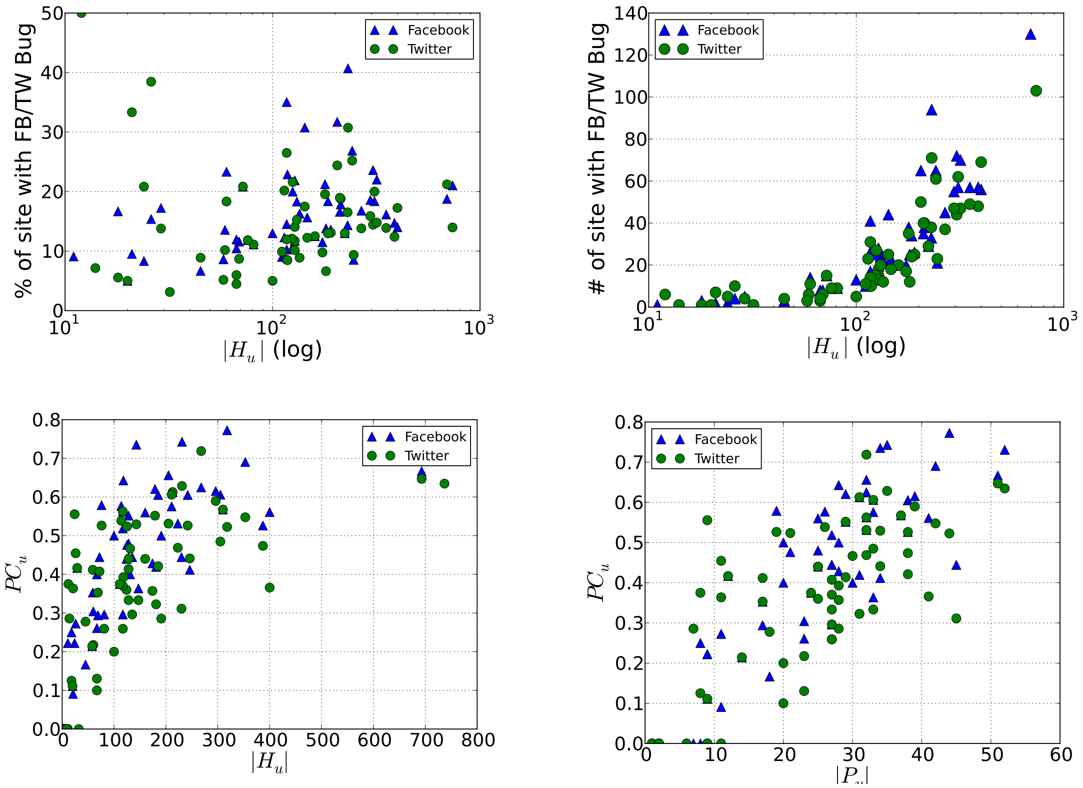

Figure 4: From left to right (a) Histrory size VS History coverage(%); (b) Histrory size VS History coverage(#); (c) History size Vs Profile coverage; and, (d) Profile size Vs Profile coverage.

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 5a 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.

4.3.2 User Web History Analysis

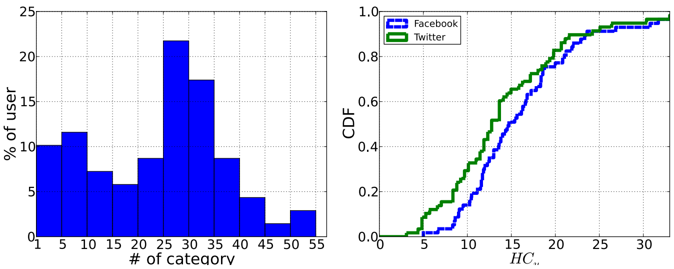

Figure 5: Profile length distribution (left),CDF of history coverage (right)

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 4c, 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 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.

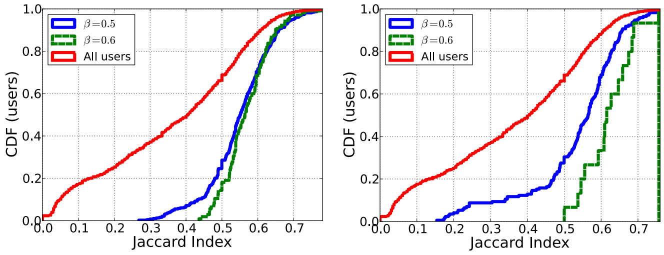

Figure 6: P_u similarity using Jaccard index (FB left, TW right)

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 β 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 computed as $J(i,j) = \frac{|P_i \cap Pj|}{|P_i \cup P_j|}$. For example, a Jaccard value $J(i,j)$ of 0.5 means that user i and user j share the half of their respective profiles P_i and P_i. Figure 6 shows the CDF of the Jaccard index for different values of β (i.e., 0.5, 0.6 and 0). First, the red curve depicts Jaccard values for all users: we observe that nearly 80% of the users have a similarity value lower than 0.5 which suggests a weak correlation between user profiles. However, blue ($\beta = 0.5$) and green ($\beta = 0.6$) curves show that there is a high correlation between users who are highly tracked. For instance, all users who have PC_u higher than 0.6 for Twitter have a Jaccard index above 0.5. This means that all these users share at least 50% of their profiles. This clearly indicates that highly tracked users tend to have similar profiles.

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`, 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

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 rise 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

[1] Alerts users about the web bugs, ad networks and widgets on visited web pages. www.ghostery.com.

[2] Facebook Sets Historic IPO. online.wsj.com/article/SB10001424052970204879004577110780078310366.html.

[3] Method and system for web user profiling and selective content delivery. http://www.google.com/patents/US8108245.

[4] TrustedSource - Customer URL Ticketing System. www.trustedsource.org/en/feedback/url.

[5] S. M. Bae, S. H. Ha, and S. C. Park. Fuzzy web ad selector based on web usage mining. *IEEE Intelligent Systems*, 18:62–69, 2003.

[6] L. Humphreys, P. Gill, and B. Krishnamurthy. Privacy on twitter: How much is too much? privacy issues on twitter. *the annual meeting of the International Communication Association*, 2010.

[7] B. Krishnamurthy, K. Naryshkin, and C. E. Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Web 2.0 Security and Privacy Workshop*, 2011.

[8] B. Krishnamurthy and C. Wills. Privacy diffusion on the web: a longitudinal perspective. In *WWW '09: Proceedings of the 18th international conference on World wide web*. ACM, 2009.

[9] B. Krishnamurthy and C. E. Wills. Privacy leakage in mobile online social networks. In *Proceedings of the 3rd conference on Online social networks*, WOSN'10, 2010.

[10] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the 11th ACM/USENIX Internet Measurement Conference (IMC'11)*, Berlin, Germany, November 2011.

[11] A. Roosendaal. Facebook Tracks and Traces Everyone: Like This! *SSRN eLibrary*, 2010.

[12] T.-F. Yen, Y. Xie, F. Yu, R. P. Yu, and M. Abadi. Host fingerprinting and tracking on the web: Privacy and security implications. In *Proceedings of the 19th Annual Network & Distributed System Security Symposium (NDSS)*, Feb. 2012.