Privacy Scoring and Users' Awareness for Web Tracking

Asma Hamed, Hella Kaffel-Ben Ayed
CRISTAL Lab.
ENSI, University of Manouba
Tunis, Tunisia

Hella Kaffel-Ben Ayed
Faculty of Science of Tunis
University of Tunis El Manar
Tunis, Tunisia

***Abstract*—Web tracking raises a potential privacy concern since users' browsing activities and personal information are collected and revealed to third parties without the users' consent. The emergence of new pervasive systems offers opportunities to widen "traditional" Web to the world of pervasive devices. In this paper, we revisit the problem of Web tracking. We propose an intuitive scoring model to measure the users' privacy risk when they are browsing the Web. We present a Firefox add-on which computes the scores of visited Webpages and provides users' awareness about how there are tracked on the Web. We distributed this add-on to two sets of Tunisian volunteers (students and researchers). We analyze the statistical results of the collected datasets and show that the users' behavior, the trackers and tracking components can increase the risk of privacy violations on the Web.**

Keywords— Web tracking, tracking components, privacy, scoring model.

I. INTRODUCTION

Tracking is the act or process of following something or someone [1]. Tracking aims at collecting information about users' behaviors, activities and location. Web tracking permits to identify the visitors of Websites as well as the visited Webpages, in order to know the visitors' interests and behaviors and to provide them with personalized content [2]. Web tracking is usually implemented by advertising companies to customize offers to suit customer's preferences. Web tracking is essential to advertising which in turn supports the Internet ecosystems. However, this raises a potential privacy concern if data related to users' activities and personal information is revealed to third parties without their awareness. In both desktop context and mobile context, Web tracking is implemented using various tracking components, such as cookies, JavaScripts (JS), Local shared objects (LSOs) and Iframes, and relies on third party trackers. Cookies are small pieces of text sent from a server and stored in a client computer in order to register the user's browsing state [3]. LSOs are a new generation of persistent cookies that has spread across the Web [4]. An iFrame (Inline Frame) element is often used to embed contents within the current Webpage such as an advertisement sent from a third party. Iframes permit the collection of information related to users without their consent [5]. JavaScript (JS) can be used to send, update and modify cookies as well as to manage the interaction between cookies and LSO or simply launch iframes. All these tracking components permit users' identification, observation, monitoring as well as the leakage of their browsing activities and personal information to third parties which constitutes an issue to informational privacy. The emergence of new pervasive networks and devices with novel capabilities, offers opportunities to widen "traditional" Web from the world of computers to the world of mobile and pervasive devices. Pervasive Web has to cope with privacy threats resulting from Web tracking.

Researchers have conducted studies to measure tracking. They also stressed the risk of privacy violations resulting from Web tracking. Tools have been proposed to detect and quantify tracking over mobile devices either during browsing or using Website mobile applications. However, to the best of our knowledge there is no approach that addresses users' awareness by identifying tracking parameters associations between parameters and privacy scoring. Our goal in this paper is to address this issue. We propose an approach to improve users' awareness based on scoring the privacy risk and based on identifying the parameters as well as associations between parameters that constitute a high risk to users' privacy. We propose an intuitive scoring model to measure to which extent users' privacy is at risk when they are browsing the Web. We present a Firefox add-on that measures and alerts users about Web tracking. Two versions of this add-on (one for the desktop context and the other for mobile android devices) were distributed to two sets of Tunisian volunteers (students and researchers). We proceed to an empirical study to better understand which trackers act and how they act and to analyze the effect of tracking findings of interest regarding privacy considerations.

The remainder of this paper is organized as follows: Section II presents related works on this topic. Section III presents the scoring model. Section IV describes our add-on. Section V depicts the statistical analysis of our Web tracking study. Finally,

978-1-4799-7349-1/15/$31.00 ©2015 IEEE

in Section VI we briefly conclude the paper.

II. RELATED WORK

Various studies aim to alert internet users by developing tools in order to emphasize privacy-violating information flows on the Web. Self-Destructing Cookies is a Firefox add-on which gets rid of cookies as soon as users close their tabs. Tracking cookies are detected and removed immediately [6]. No Google Analytics is also a Firefox add-on designed for a specific and popular tracker: Google Analytics. It implements a content policy which blocks all network requests within Firefox for anything from the Googleanalytics.com domain. This means JavaScript, gif images, and everything else is blocked from that domain [7]. Firefox add-on Ghostery is very popular for many desktop browsers and it has recently released a mobile version for android. Ghostery detects trackers and offers the possibility of blocking them or being selective in blocking. Ghostery has a large database of the trackers components signatures and this database continues on growing with every update of the add-on. It presents as an option the possibility of sending information about trackers from the user browser to Ghostery servers [8]. Such as Ghostery, Privacy Defense detects trackers and offers the possibility of customized blocking. The operation is based on finding trackers' signatures. It is available for downloading on Mozilla add-ons site, but it's not fully operational before paying for it. This can have a great impact on this add-on use, since there are other free available add-ons [9]. Privacyscore add-on for Firefox and Chrome is a project of PrivacyChoice, which was founded in 2009 to make privacy easier for Websites, apps and their users by detecting trackers present in the visited Websites [10]. Privacyscore audits only the most famous and the most visited Websites. Lightbeam add-on creates a real-time graph of all the tracking cookies being deposited on users' browser as they move around the Web. One of the inconvenient of lightbeam is that after visiting about four or five sites, the graph tends to get really confusing and it's hard to tell which advertisers are connected with which sites. Besides lightbeam relies only on cookies, it does not cover other tracking components such as JS, iframes, LSOs, etc.

Other studies have addressed the issue related to third party trackers. Authors in [11] study the prevalence of third-party trackers on the Web. They explored more than 1200 popular Websites and show that the penetration of the top-10 third-party tracking servers viewing user habits has increased from 40% in Oct'05 to 70% in Sep'08. The same authors demonstrate in another work [12] that personally identifiable information belonging to any user, such as name, gender or OSN (Online Social Network) unique ID, is leaked to third-party servers via the OSN. Thus, besides viewing the surfing habit of some users, third parties associate also the habits with a specific habit and potentially gather much more personal information. This ability to link information across Websites and OSN raises important privacy concerns. In [13] authors show that most commercial pages are tracked by multiple parties. Trackers vary widely in their coverage with a small number being largely deployed while others rely on combinations of tracking behaviors. Based on Web search traces taken from AOL data, they also show that each tracker can capture more than 20% of a user's browsing behavior. In [14] authors study the tracking in the most popular Websites through different countries. They highlight the dominance of US trackers in China and Europe while in Russia they observe the dominance of local trackers.

As for tracking in mobile devices, in [15] authors compare tracking on desktop and mobile devices. They show great similarities between tracking in both platforms, new trackers were barely found. A closer look to tracking components reveals that the average presence of cookies is more important on the desktop than on the mobile devices, which is probably due to the limited local storage on mobile devices. Authors show in [16] that MOSNs (Mobile Online Social Networks) leak private information to third parties, such as user's precise location, his gender or name, and even subject's unique social networking identifier. This type of leakage can be used to determine the identity and to build an accurate profile of MOSN users. In [17] authors underline the unawareness of users regarding geolocation information contained in the photos and videos taken with their smart phones or cameras. Such information may be used for location tracking or aggregated with data collected from OSN. This may be considered as a potential source of information leakage and may lead to a privacy violation.

III. THE PRIVACY SCORING MODEL

We propose a scoring model to quantify the risk to privacy resulting from the use of the previously presented components. For each visited Webpage we compute a privacy score called total score. Our scoring model is computed intuitively based on the assumptions that the presence of a tracking component is a risk by itself. Moreover, the presence of two components, mainly JS and one of the other components, increases the risk since the interaction between them might permit to collect more data about users [18]. We define an existence score which is incremented each time the component is detected and an interaction score which is incremented whenever a tracking component and a JS are present in the same Webpage. The total score of a Webpage is computed as the sum of existence score and interaction score as shown in (1):

$$Total\_Score = Existence\_Score + Interaction\_Score \quad (1)$$

TABLE I summarizes and describes the different scores.

TABLE I. Notation table

Definition	Score
measure the presence of tracking components	Existence_Score
measure the presence of LSO	LSO_Score
measure the presence of cookies	Cookie_Score
measure the presence of iframes	Iframe_Score
measure the presence of JS	JS_Score
measure the possible interaction between JS and the tracking components	Interaction_Score
measure the possible interaction between JS and cookies (JS/Cookie)	JS/Cookie_Score
measure the possible interaction between JS and LSO (JS/LSO)	JS/LSO_Score
measure the possible interaction between JS and iframes (JS/Iframe)	JS/Iframe_Score
measure the ultimate score as the sum of Existence_Score and Interaction_Score	Total_Score

A. *The existence score*

This score aims to measure the presence of tracking components. It is the sum of the scores of these components as shown in (2):

Existence_Score = LSO_Score + Iframe_Score + JS_Score + Cookie_Score (2)

In the following we depict the way we compute the score of each component.

LSO and cookie Scores: Both of cookies and LSOs have a Time To Live (TTL). Cookies TTL varies from few minutes to several years. LSOs are persistent and have unlimited TTL which might increase privacy risk. We consider that the higher TTL is the higher is the risk of privacy leakage. We assume here that LSOs and cookies TTL_MAX = 10 years. We measure the privacy risk as TTL/TTL_MAX. So for each detected cookie or LSO, the score is incremented by: 1 + TTL/TTL_MAX. For LSOs, since TTL=TTL_MAX, the score is incremented by 2.

Iframe and JS Scores: Unlike cookies and LSOs, iframes and JS do not admit a TTL since they do not persist on the client side. We assign 1 point in the score to measure the presence of an iframe or JS.

B. *The interaction score*

The concept of interaction may be defined as the combination of JS and one of the previously identified tracking components within a same Webpage. A JS code can be used to send, update or modify a cookie. It ensures the interaction between cookies and LSOs triggers an iframe [19]. This interaction might increase the privacy threat. Thereby we define an interaction score that complements the existence score. Interaction-Score is computed as follows:

Interaction_Score = JS/Cookie_Score + JS/Iframe_Score + JS/LSO_Score (3)

Sending the content of a cookie by a JS may imply a dangerous activity of information disclosure contained in the cookie. This activity increases the JS/Cookie_Score by 1. In the same way, we increment the JS/Iframe_Score by 1 since a JS may be able to launch an iframe or to send information via it as we mentioned previously. On the other hand, we consider that updating, changing information in the cookie or regenerating a cookie using a LSO is similar to recreating it (new TTL, new data, ...), hence we add the Cookie-Score to both JS/LSO-Score and JS/Cookie-Score.

IV. THE PRIVACY ADD-ON

Making users' aware about how they are tracked constitutes a first step to privacy provision. We developed two versions of a Firefox add-on, named TrackScore (the desktop version) and TrackscoreMobile (the android version).They detect the tracking components used by trackers and computes, for each Webpage and each detected tracker, the corresponding score. This process is done locally within the users' device. The results are sent to a server side database. A small TrackScore icon is visible on the top right of the browser's window and displays the score of the visited Webpage and the list of trackers detected on this page. This design keeps the user informed about his scores and about the trackers without disturbing his activity. Optionally, the user can visualize the different tracking components for each tracker by clicking on the icon. The icon color changes with the computed score i.e. the relative level of tracking threats. A red icon alerts that the visited Webpage exhibits a high risk, i.e. the computed score is relatively high. A yellow icon is for moderately high score and a green icon is for a low tracking score. By selecting the "Statistics" option, the user gets a summary of the assessment of the visited Webpage. This summary contains: (1) the score of the visited Webpage, (2) the average score of the visited Webpage for the specific user, (3) the average score of the visited Webpage for all users, (4) the user's average score of all visited Webpages, (5) the average score of all users and (6) the detected tracking components on the visited Webpage. As For TrackScoreMobile, it displays the score of the visited Webpage on the bottom of the screen of the mobile device. Optionally, the user can display the tracking components for each tracker by clicking on "More information about trackers". He/she can also find more details about the tool by clicking on "About TrackScore Mobile". Fig. 1 presents an example of TrackScore's and TrackScoreMobile's screen shots.

(a) TrackScore screen shot

(b) TrackScoreMobile screen shot

Fig. 1. The privacy add-on

V. WEB TRACKING EXPERIMENT

Our goal is to address these questions: (1) what is the relationship between the users' behavior and the potential tracking, (2) what are the most dangerous trackers, (3) what are the most used tracking components. We consider three metrics to measure tracking: (1) number of detected trackers, (2) number of tracking components and (3) computed score.

A. TrackScore experiment

We sent TrackScore by email to a set of volunteers, mainly Tunisian students and researchers. The experiment was conducted among 14 users from December 2012 to June 2013. We collected 3849 observations: 184 Websites, 15 domains and 583 trackers.

Users' behavior vs. users' tracking

We focus first on the users' lifetime duration, i.e. for how long the add-ons have been used during the data collection period, in order to select the users that have a significant activity for our study and study the potential impact of users' behavior on their tracking. The average users' lifetime is 4.35 days. Users 1 (9.36% of the total collection's duration), 2 (6.40% of the total collection's duration), 3 and 4 (4.43% of the total collection's duration) have the most significant life-time duration. Observations show that users have mainly visited .com, .fr, .edu, .net and .tn domains. However foreign domains are more visited than the .tn. This can be explained by the small number of national domains which provide the requested services to this kind of population (students and researchers).

The mostly tracked users are 1 (48.20% of the trackers), 3(48.20 of the trackers), 4(23.84% of the trackers) and 2(19.55% of the trackers. We observe that users having mostly used the add-on are the mostly tracked. Based on these observations, we can conclude that the number of trackers targeting a user increases with the browsing period of time.

As for the average score for each user, we observe that the most tracked users are not the ones with the highest average score. This may let us assume that the simple fact of browsing exposes users to tracking components independently of the browsing time duration.

Trackers' statistics

The issue here is to identify the most frequent trackers and how much they rely on tracking components in order to identify the parameters that are the most dangerous for Web users' privacy. For that, we focus on: (1) the top 10 trackers in the visited domains and the visited Websites, (2) the top 10 trackers in terms of used tracking components, (3) the means of the different tracking components for the top 10 trackers and (4) the top 10 trackers in terms of score.

We observed that Google and Facebook take the first places in the experiment. This can be explained, as found in related works, by the fact that Google relies on several third party trackers: Google-analytics.com, gooleapis.com, Googlesyndication.com and doubleclick.net. Facebook relies on Facebook.com and Facebook.net. Then we focused on the how trackers rely on tracking components. We found that unlike the previous results, Google.com and Facebook.com are not even among the top 10 trackers in terms of number of tracking components. We explored the means of the different tracking components for the top 10 trackers in terms of tracking components. We observed that cookies, JS and JS/Cookie are the most used components by the top 10 trackers in terms of tracking components. Finally we found a similarity between the top 10 trackers in terms score and the top 10 trackers in terms of tracking components. In order to confirm this observation, we computed the Jaccard index. We found that the computed scores and the number of used tracking components are similar in 81.81% of the cases. This indicates a prospective dependency between the scores and the tracking components and can validate qualitatively our proposed scoring model.

B. TrackSoreMobile experiment

The experiment gives statistics on trackers, tracking components and scores in order to understand and measure Web tracking during spontaneous browsing. We sent TrackScoreMobile by email to a set of volunteers, mainly Tunisian students and researchers. The experiment was conducted among 7 users from September 2013 to November 2013. We collected 22109 observations: 229 Websites, 20 domains and 497 trackers. The collected tracking information is studied and the results will be exposed in the following. We focus

first on the users' lifetime duration, i.e. for how long the add-ons have been used during the data collection period, in order to study the potential impact of users' behavior on their tracking. The average users' lifetime is 33.28 days. Users 1 (91.66% of the total collection's duration), 2 (77.46% of the total collection's duration), 3 (66.66% of the total collection's duration), and 4 (63.88% of the total collection's duration) have the most significant lifetime duration. The other users have less than 20% of the total duration. Observations show that users have mainly visited .com, .fr, .edu, .net and .tn domains. Foreign domains more visited than the .tn. This can be explained by the small number of national domains which provide the requested services to this kind of population (students and researchers). We computed the trackers' percentage of each user. We found that the mostly tracked users are 4 (62.37% of the trackers), 2 (39.83% of the trackers), 1(30.38% of the trackers) and 3(10.86% of the trackers). We observe that users having mostly used the add-on are the mostly tracked. Based on these observations, we can conclude that the number of trackers targeting a user increases with the browsing period of time.

As for average scores, users with the highest scores are 4 (46.37 as average score), 2 (36.91 as average score), 1 (30.66 as average score) and 3 (30.61 as average score). The other users' average scores is less than 20. We observe that the most tracked users are also the ones with the highest average score. To confirm this observation we find a strong correlation between the score and the lifetime. Pearson Coefficient is 0.862 with a p-value 0.013.

To identify the most frequent trackers and how much they rely on tracking components, we found that that Google and Facebook take the first places in the two experiments. Google relies on several third party trackers: Google-analytics.com, gooleapis.com, Googlesyndication.com and doubleclick.net. Facebook relies on Facebook.com and Facebook.net. We also observe that cookies, JS and the combination of JS and Cookie are the mostly used tracking components. Regarding the average number of tracking components per tracker, we found that trackers with highest average number of tracking components are cdiscount, Quinn street, m.achetezfacile and mobile. clubic. The most frequent trackers are not those with highest number of tracking components.

To summarize our findings here, results of the collected datasets analysis show that users' tracking is impacted by users' behavior since the number of trackers targeting a user as well as the score increase with the browsing period of time for both data collections. We also observed that the most frequent trackers are not the ones which use the largest number of tracking components. An issue can be identified here: what is the most risky for the user's privacy between the presence of a given tracker and the number of tracking components it uses? Let's take Facebook as an example: It is present in most of the visited domains and Websites but it does not use a large number of tracking components. However, when a user visits a Webpage containing Facebook plugins (e.g like, share, etc.) the collected data can be linked to his Facebook account with his personal information. Regarding this issue, we consider that the frequency of a tracker might be more risky for the user's privacy than the number of tracking components it uses.

VI. CONCLUSION

In this paper, we defined an intuitive scoring model to measure to the users' privacy risk when they are browsing the Web. We developed two versions of a Firefox add-on: TrackScore for desktop and TrackScoreMobile for Android devices. This add-on computes the score and alert users' about trackers and scores. We distributed the two versions of the add-on to two sets of Tunisian students and researchers and collected scores and tracking information in a server side application.

The first results of the collected datasets analysis show that users' tracking is impacted by users' behavior since the number of trackers targeting a user increases with the browsing period of time for both data collections in desktop and mobile wild experiment. The score of desktop users is not dependent on the browsing lifetime while for mobile wild users the scores increase with the browsing durations. This result is a high risk to mobile users' privacy since they are permanently connected to the Web. Results show also great similarities between trackers in desktop and mobile wild platforms. New trackers were barely found. Google and Facebook are the most frequent trackers. However they are not the ones using the largest number of tracking components. We can identify here an issue: what is the most risky for the user's privacy between the presence of a given tracker and the number of tracking components it uses? Let's take Facebook as an example: It is present in most of the visited domains and Websites but it does not use a large number of tracking components. However, when a user visits a Webpage containing Facebook plugins (e.g like, share, etc.) the collected data can be linked to his Facebook account with his personal information. Regarding this issue, we consider that the frequency of a tracker might be more risky for the user's privacy than the number of tracking components it uses. The more a tracker is present the more is the amount of collected data, which increases sensitive information, linked to users' activity resulting in the creation of full and accurate browsing profiles. As users become more and more dependent on social networks and other popular Websites, the trackers become more knowledgeable about users' browsing behavior and habits.

REFERENCES

[1] A. Bhaduri, "User controlled Privacy protection in location-based services", Master of science thesis in Spatial Information Science and Engineering, 2003, Chap 3, pp 32.

[2] S. Mlot, "Parents Worried About WebFirms Tracking Their Kids", http://www.pcmag.com/article2/0,2817,2412313,00.asp, 2012, access date: March 2014.

[3] S. Mittal, "User privacy and the evolution of third-party tracking mechanisms on the world wide web", Ph.D. dissertation, Stanford University, 2010.

[4] A. Soltani, S. Canty, Q. Mayo, L. Thomas, C.J. Hoofnagle, "Flash cookies and privacy," http://papers.ssrn.com/sol3/papers.cfm?abstract-id=1446862, 2009, access date: September 2012.

[5] M. Rouse, "Betterprivacy",http://whatis.techtarget.com/definition/IFrame-Inline-Frame, 2011, access date: September 2012.

[6] Ove, "Self-destructing cookies", https://addons.mozilla.org/fr/android/addon/self-destructing-cookies/?src=search/,2013, access date: March 2014.

[7] E. Vold, "No google analytics", https://addons.mozilla.org/fr/android/addon/no-google-analytics/?src=search, 2013, access date: March 2014.

[8] J.Signanini and F. Shnir, "Ghostery," https://addons.mozilla.org/fr/android/addon/ghostery/?src=search, 2013, access date: March 2014.

[9] "[Privacy] defense", https://addons.mozilla.org/fr/android/addon/privacy-defense/?src=search, 2013, access date: March 2014.

[10] R. Waugh, "Watch them watching you: Privacy site that gives instant rating for how websites use and abuse your details",http://www.dailymail.co.uk/sciencetech/article-2100958/Who-watching-Privacy-Score-lets-EVERY-company-watching-you-visit-webs-sites.html, 2012, access date: March 2014.

[11] B. Krishnamurthy and C. Willis, "Privacy diffusion on the web: a longitudinal perspective", in Proceedings of the 18th international conference on World Wide Web, 2009, pp. 541–550.

[12] B. Krishnamurthy and C. Wills, "On the leakage of personally identifiable information via online social networks", in Proceedings of the 2nd ACM Sigcomm Workshop on Online Social Networks (WOSN), Barcelona, Spain, August 2009, pp. 7–12.

[13] F. Roesner, T. Kohno and D. Wetherall, "Detecting and defending against third-party tracking on the web", in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012, pp. 12–12.

[14] C. Castelluccia,, S. Grumbach and L. Olejnik, "Data harvesting 2.0: from the visible to the invisible web", in Proceedings of the 12th Workshop on the Economics of Information Security, Georgetown University, Washington, D.C, June 2013.

[15] C. Eubank, M. Melara, D. Perez-Botero and A. Narayanan , "Shining the floodlights on mobile Webtracking - a privacy survey", in Proceedings of IEEE Symposium on Security and Privacy Workshops, Web2.0 Security and Privacy (W2SP), May 2013.

[16] B. Krishnamurthy and C. Wills., "Privacy leakage in mobile online social networks", in Proceedings of the 3rd Workshop on Online social networks (WOSN), Boston, MA, USA, June 2010.

[17] R. Friedland and G. Sommer., "Cybercasing the joint: On the privacy implication of geotagging", in Proceedings of the 5th USENIX Workshop on Hot topics in security, (HotSec), Washington, DC., USA, August 2010.

[18] A. Hamed, H. Kaffel-Ben Ayed, M.A. Kaafar and A. Kharraz, "Evaluation of third party tracking on the web", in Proceedings of the 8th International Conference for Internet Technology and Secured Transactions (ICITST), London, UK, December 2013.

[19] F. Roesner, T. Kohno and D. Wetherall, "Detecting and defending against third-party tracking on the Web," in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012.