Towards Privacy-Friendly Online Advertising

Julien Freudiger, Nevena Vratonjic and Jean-Pierre Hubaux
EPFL, Switzerland
firstname.lastname@epfl.ch

***Abstract*— Modern web sites commonly interact with third-party domains to integrate advertisements and generate revenue from them. To improve the relevance of advertisements, online advertisers track user activities online with third-party cookies. However, excessive online tracking might cause unreasonable access to users' browsing information. Users are thus in need of a simple way to control the sharing of their browsing information with advertisers in order to protect their privacy. We survey current techniques to conceal browsing information from third parties (e.g., block third-party cookies) and propose a novel approach that enables advertisements to have discrimination capabilities without allowing for excessive tracking of users. Our solution uses a collection of third-party cookies to restrict the tracking on a per web site basis. We present various implementations of our proposal and provide a proof of concept code to demonstrate its feasibility.**

I. Introduction

Online advertising is at the center of the Internet economy [31]. It is a large and successful business because: (i) It offers *immediate* publishing of advertisements not limited by geography or time, and (ii) it can be *personalized* by tracking users *spatially* over different web sites and over *time*. The tracking is done by exploiting client-side browser state (e.g., third-party cookies). It permits advertisers to relate advertisements to users' interest [33] and to users' online behavior [12], [20], [24]. Many advertisers are thus attracted to this new advertising distribution channel.

Web sites also benefit from hosting online advertisements as it generates revenue. In the recent years, a novel business model based on online advertising created new opportunities online for bloggers, newspapers, and web applications. Users also benefit from online advertising because it sponsors the free access to valuable content and services [33]. For example, newspapers offer articles online for free and generate revenue from the accompanying advertisements. Similarly, Google provides a competitive email service for free, and embeds advertisements with emails to sponsor its service. Finally, users also appreciate online advertising as it can provide insightful links, especially if it is well targeted [21].

However, the proliferation of online advertisements raises privacy concerns. By tracking users on the Internet, advertisers can expose their personal activities and obtain information such as consulted web pages and social network connections. For example, third-party cookies (i.e., cookies used with a third-party server of the visited web site) enable advertisers to track users across web sites affiliated with them. Hence, excessive online tracking might allow for the identification of users online [28], [29], [30].

As a consequence, several applications have emerged to limit the privacy footprint of users online by automatically blocking cookies [1], [6], [7].[1] However, blocking all first-party cookies (i.e., cookies of the visited web site) has adverse effects on surfing the web and might affect the usability of web pages. To solve the privacy/usability trade-off of first-party cookies, Shankar and Karlof [34] propose to improve the management of first-party cookies by letting users decide which cookies to block/accept based on a visual comparison of web pages with and without first-party cookies.

Similarly, blocking all third-party cookies presents a significant problem for the online advertising industry. All visits to an advertiser are still recorded, but a person who has deleted his third-party cookies is not recognized as the same returning visitor. Consequently, blocking third-party cookies makes advertising less relevant as it will be based only on the current page browsed by the user (i.e., context) and not on what the user might have done in the past (i.e., behavior). The current management of third-party cookies does not permit for the tuning of behavioral tracking done by advertisers: It allows advertisers to track users either across all web sites or none.

This paper proposes a novel solution to solve the privacy/traceability trade-off of third-party cookies (Fig. 1). It manages all cookies used with third-parties in a privacy-friendly manner. Our solution enables advertising to have differentiation capabilities without allowing for excessive tracking of users online. To do so, we assume that: (i) Advertisers want users to click on advertisements, and need to track users to improve online advertising relevance; (ii) users are willing to share some information with advertisers in order to get relevant advertisements, free content and free services.

We give users a fine-grained control of the dissemination of their information to advertisers on a *per web site* basis. To do so, our solution maintains a collection of alternative third-party cookies with each online advertiser. Third-party cookies are sent to the advertiser depending on the consulted web site. The same third-party cookie can be sent to an advertiser for different web sites if it improves the advertisements relevance without allowing for excessive tracking. The decision to share a third-party cookie across different web sites depends on the visited web site and user privacy preferences. Users can set their preferences either manually or automatically by relying on online communities [23]. We test the feasibility of our solution by implementing a Firefox extension and show that

[1] Note that instead of blocking cookies, users can also directly *opt out* from advertising on advertisers' web sites [2].

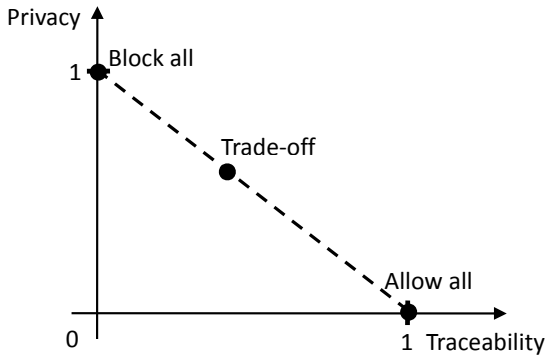

Fig. 1. Privacy/traceability trade-off. At $(1, 0)$, the default third-party cookies management allows for the complete tracking of users online. At $(0, 1)$, blocking all third-party cookies impedes online tracking by third parties. Our solution allows to trade-off privacy and traceability by limiting spatial and temporal traceability of users online.

users' traceability can be controlled without requiring any changes from advertisers. This paper is part of the recent trend of providing tools to help individuals reduce their privacy footprint online [13], [14], [15], [16], [34].

II. Preliminaries

A. HTTP Cookies

HTTP Cookies[2] are data items stored in the user browser, which are assigned to users by web servers. On subsequent visits, browsers send back the cookies to web servers, along with HTTP requests. Cookies that are sent to the server hosting the visited web page are called *first-party cookies* (FP-cookies). FP-cookies are used by web servers to keep the state of the connection, e.g., to differentiate users. As web pages might contain references to components needed to render the page (e.g., images or advertisements), web browsers issue additional HTTP requests for these elements. If the elements are stored on servers in other domains, cookies that are sent during the retrieval of these components are called *third-party cookies* (TP-cookies). TP-cookies allow third-party servers to track users across websites. In practice, a cookie can be used as first-party cookie or third-party cookie (e.g., a website can operate both in first-party and third-party mode). Hence, in this paper, we consider the privacy-friendly management of all cookies sent to third-parties.

Cookies are usually set with the *Set-Cookie* HTTP header and sent with the *Cookie* HTTP header. The *Set-Cookie* header is sent by the server in response to an HTTP request from a user to create a cookie in the user's browser. Cookies come in two flavors: *Session* cookies have no expiration date and expire after the Internet session ends, whereas *persistent* cookies are long-lived. For each HTTP message sent to a server, if there is a cookie in the browser that matches the server, the cookie is included by the browser in the HTTP *Cookie* header.

[2]From here on, referred to as cookies.

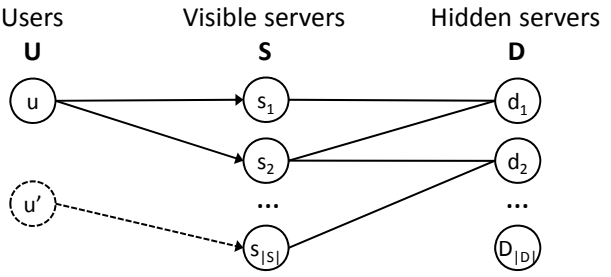

Fig. 2. Tripartite graph G. Visible servers are *associated* if they share a connection with a hidden server. Visible servers are *connected* to a hidden server if they host advertisements from the hidden server. With our solution, user u can use multiple TP-cookies to appear as two different users u and u' to the hidden server d_2.

B. System Model

In compliance with [29], we denote web servers accessed by users to download first-party components as *visible* servers and servers accessed to download third-party components as *hidden* servers. A visible server is *connected to* a hidden server when the access to the visible server causes the access to the hidden server.

We model the relation between users, visible servers and hidden servers with a tripartite graph $G = (U \cup S \cup D, E_1 \cup E_2)$ where U is the set of users, S is the set of visible servers and D is the set of hidden servers (Fig. 2). A user $u \in U$ can connect to a visible server $s_i \in S$ (a visible server is equivalent to a web domain). The web domain can host several web sites $b_\ell \in B_i$, where B_i is the set of web sites in the web domain s_i. A web site is uniquely identified with a URL. A visible server is connected to a hidden server $d_j \in D$ if it hosts content from d_j. In other words, an edge $(u, s_i) \in E_1 \subseteq U \times S$ exists if a user in U visits a visible server in S. An edge $(s_i, d_j) \in E_2 \subseteq S \times D$ exists if a web server redirects its users to a hidden server in D. We assume that the web browser of a user u keeps the history of accessed web sites $H_u(B)$, where $B = \bigcup_{s_i \in S} B_i$ and the history of accessed hidden servers, $H_u(D)$. Web browsers also store cookies $c_k \in K$, where K is the set of all cookies in the system and remember the server that caused their assignment. For example, in Tab. I, the TP-cookie $ID = agbfd12$ is related to $doubleclick.net$. We denote $H_u(K)$ the set of all the cookies stored in the browser of a user u. Without loss of generality, we focus on a single user in our analysis. Consequently, we omit the index u.

C. Online Advertising

Online advertisers track users online to improve the efficiency of online advertising:

- *Contextual tracking* allows for the real time targeting of advertisements to the content of a page (e.g., Gmail).
- *Behaviorial tracking* allows for the use of information about previously and currently browsed web pages.

A popular technique to track users online makes use of persistent TP-cookies. Online advertisers set an identifying TP-cookie in the browser, which will be sent back each time

Visible Server	Hidden Server	Third-party Cookie
www.orkut.com	doubleclick.net advertising.com	ID=agbfd12 ID=19576981
www.myspace.com	doubleclick.net	ID=agbfd12
www.sourceforge.net	doubleclick.net quantserve.com	ID=agbfd12 user=97v124ag3
www.mininova.org	quantserve.com	user=97v124ag3

TABLE I

EXAMPLE OF THIRD-PARTY COOKIES OF USER u. DOUBLECLICK.NET CAN TRACK USER u ACROSS THREE DIFFERENT WEB SITES.

Visible Server	Hidden Server	Third-party Cookie
www.orkut.com	doubleclick.net advertising.com	ID1=agbfd12 ID=19576981
www.myspace.com	doubleclick.net	ID1=agbfd12
www.sourceforge.net	doubleclick.net quantserve.com	ID2=2pokn92 user1=97v124ag3
www.mininova.org	quantserve.com	user2=012nfnaw2

TABLE II

EXAMPLE OF THIRD-PARTY COOKIES OF USER u. TP-COOKIES ARE MODIFIED TO LIMIT THE PROFILING ACROSS WEB SITES.

the browser sends a request to the advertiser together with an IP address, a URL, and a referrer. The referrer identifies, from the point of view of an Internet resource, the URL of the resource which links to it. Online advertisers can thus track users *temporally*: Multiple visits of the same user on the same web site can be identified by online advertisers. They also track users *spatially*: Users are tracked across different web sites connected to the same advertiser.

Consider the example in Tab. I: When user u browses $orkut.com$, which hosts advertisements from $doubleclick.net$, u is assigned a TP-cookie from $doubleclick.net$ during their first communication. Then, if u browses another web site also hosting advertisements from $doubleclick.net$ (e.g., $myspace.com$), u's browser will use the previously assigned TP-cookie with the HTTP packets sent to $doubleclick.net$. Therefore, $doubleclick.net$ learns that u has visited both $orkut.com$ and $myspace.com$ by checking the referrer and can track u spatially over the two visible servers.

As in [29], we say that visible servers are *associated* when they are connected to one or more common hidden servers. In Fig. 2, visible servers s_1 and s_2 are associated as they share the hidden server d_1. In order to keep track of associations between visible servers, we say that the TP-cookie c_k is linked to (s_i, d_j) if s_i is the visible server that caused the communication with the third party d_j. We also denote with $\upsilon(s_i, d_j, c_k)$ the number of visits of a user to a hidden server d_j with the TP-cookie c_k, caused by the visible server s_i.

There are other techniques to track users online. HTTP redirections for example can also be used to track users with first-party cookies. However, this technique is not as popular as tracking based on TP-cookies as shown in Section IV. In addition, Doppelganger browser extension [34] thwarts such tracking. Other tracking techniques are discussed in Section V.

D. Threat Model

As cookies are used to identify subsequent visits of users, they increasingly reveal more information about users' browsing habits (Tab. I). The traceability of users based on TP-cookies was characterized in [28], [29], [30] showing that a majority of web servers are associated with at least one other visible server. The HTTP referrer also reveals sensitive information as it identifies the visited visible server. For example in Tab. I, the referrer of user u accessing the hidden server $doubleclick.net$ from $orkut.com$ will contain the full URL of the web page browsed on the visible server, thus revealing to the advertiser the social graph of user u.

Advertisers can thus learn a significant amount of information about users' activities online. The threat is exacerbated if the collected data permits to infer users' real identities. Users' *privacy* with respect to online advertisers is thus protected if the users have the ability to prevent third parties from tracking their activities online. Note that we do not assume cooperative tracking [25], i.e., web sites do not cooperate with online advertisers to track users.

III. PRIVACY-FRIENDLY COOKIE MANAGEMENT

In order to control the information shared with advertisers, we propose to regulate the use of TP-cookies on a per web site basis depending on the visited web site (Tab. II) and on user privacy preferences. The solution is automated and allows users to control the privacy/traceabity trade-off. The decision to use a TP-cookie across different web sites connected to a same advertiser depends on the trade-off between the *benefit* caused by the TP-cookie and its associated privacy *cost* (or amount of privacy loss).

The benefit of including TP-cookies is measured by the improved relevance of the served advertisements. The privacy cost depends on the amount of information shared with advertisers. We categorize the cost in two groups, namely the spatial and temporal traceability. In this section, we propose three approaches for the privacy-friendly management of TP-cookies that differ in the achieved trade-off.

A. No Spatial Tracking across Domains and Limited Temporal Tracking

A simple approach to limit the privacy cost consists in completely preventing spatial tracking across domains and only allowing for limited temporal tracking: TP-cookies can be used for a certain period of time L_T or for at most L_V visits to the same web domain. The TP-cookie management policies are:

- **Spatial tracking policy:** For each new web domain s_i, connected to a known hidden server $d_j \in H(D)$, existing TP-cookies (if any) are not sent and instead a new TP-cookie is assigned by the third-party d_j.
- **Temporal tracking policy:** For a known web domain s_i, the same TP-cookie $c_k \in H(K)$ is used with requests to the third-party server d_j for the time period L_T or for at most L_V visits, $\upsilon(s_i, d_j, c_k) < L_V$.

This approach allows users to minimize the privacy cost: No spatial tracking is allowed except within a web domain for a limited time period. However, it also reduces the potential benefits of online advertisements because no information is shared with third parties across domains. In summary, this approach limits spatial tracking to L_S web sites, where L_S is the maximum number of web sites hosted by a web domain, and temporal tracking to L_V or L_T.

B. Limited Spatial and Temporal Tracking

To improve the relevance of advertisements, in the second approach, users share information with a limited number of associated web sites. To keep the privacy cost acceptable for users, the spatial tracking is limited to at most L_S web sites per *category* $\mathcal{C}$. Categories determine the type of web sites (e.g. business, news), hence limiting the tracking of online advertisers to specific topics. We rely on the existing categorizations of web sites based on URLs [3], [9], [12]. We assume that there is a fixed number of categories N_C and that each web site belongs to a single category. The TP-cookie management policies are refined such that, for each category, a TP-cookie can be sent for at most L_S web sites:

- **Spatial tracking policy:** Each new web site $b_\ell \notin H(B)$, connected to a known hidden server $d_j \in H(D)$, is automatically classified into one of the N_C categories. If b_ℓ belongs to category $\mathcal{C}$ and if there is a TP-cookie c_k assigned by d_j to web sites in the category $\mathcal{C}$, we verify before using c_k that:

$$\sum_{b_m \in H(B) \cap \mathcal{C}} \beta(b_m, d_j, c_k) < L_S \tag{1}$$

  where $b_m \in s_m$ and

$$\beta(b_m, d_j, c_k) = \begin{cases} 1 & \text{if } c_k \text{ is linked to } (s_m, d_j) \\ 0 & \text{otherwise.} \end{cases}$$

  In other words, if the number of times the cookie c_k was used with the third-party d_j in the category $\mathcal{C}$ is under the limit L_S, then the TP-cookie c_k can be associated with requests to the pair (s_m, d_j). Otherwise, c_k is not sent and a new TP-cookie is assigned by the third-party d_j.
- **Temporal tracking policy:** An existing TP-cookie $c_k \in H(K)$ can be used with a known web site $b_\ell \in H(B)$ in category $\mathcal{C}$ connected to the third-party server $d_j \in H(D)$ for the time period L_T or for at most L_V visits, i.e.:

$$\sum_{b_m \in H(B) \cap \mathcal{C}} \upsilon(b_m, d_j, c_k) < L_V \tag{2}$$

Consider the example in Tab. II with $L_S = 5$. The third-party $doubleclick.net$ can track user u on $orkut.com$ and $myspace.com$ because the same TP-cookie $ID = agbfd12$ is sent for both web sites. In this case, the TP-cookie was shared because both web sites belong to the same category (social networks) and the threshold is $L_S > 2$. However, the TP-cookie was not shared with $sourceforge.net$ (different category). Hence, user u appears as a different user u' to the advertiser (Fig.2).

With these policies, third-parties can profile users on a limited number of associated web sites of the same category and only during a limited time period. Hence, they can target advertisements to specific categories and improve the relevance of advertising for those. This approach has two drawbacks: (i) The number of web sites over which users can be tracked in each category is fixed, and (ii) all web sites are treated equally, as if they revealed the same information to third parties.

C. Weighted Spatial and Temporal Tracking

In this approach, we attribute weights to web sites based on two criteria: First, certain web site categories induce a higher privacy cost on users, whereas others bring more value to advertisers [8]. Second, URLs leak information depending on their content and their length. Hence, we propose to weigh web sites based on their category and the specificities of their URLs.

Web site categories: Individual users perceive differently the value of their browsing information and the potential privacy costs. Hence, the decision to reveal interest in certain web site categories should be based on user privacy preferences. We model users' preferences by assigning a weight $\omega_1(b_\ell) \in [1, N_C]$ to each web site depending on its category. The granularity of users' preferences depends on the number of categories N_C. If a category is assigned a high weight, it means that it contains sensitive information that should not be shared with third parties. For example, social networks category can be assigned a higher weight than shopping web sites.

URL specificities: URLs that contain information specifically identifying user activities are more valuable to advertisers than generic URLs. For example, *www.google.ch/search?hl=en&q=computers* reveals the user's interest in computers, his preferred language (English) and his probable location (Switzerland); thus it is more valuable than *www.google.com*. The privacy cost of a URL can be computed with regular expressions by comparing strings in the URL with predefined n-grams[3] (e.g., *q=*, *hl=*) [11]. Each URL is thus evaluated on-the-fly and assigned a weight $\omega_2(b_\ell) \in (0, 1]$. If the weight of a URL is high, then it means that it contains potentially sensitive information.

The total privacy cost $\gamma(b_\ell, d_j, c_k)$ of visiting a new web site $b_\ell \notin H(B)$ with a TP-cookie c_k associated with d_j is a weighted product of privacy costs based on the two criteria:

$$\gamma(b_\ell, d_j, c_k) = \begin{cases} \frac{\omega_1(b_\ell) \cdot \omega_2(b_\ell)}{N_C} & \text{if } c_k \text{ is linked to } d_j \\ 0 & \text{otherwise.} \end{cases}$$

where the number of categories N_C normalizes the cost $w_1(b_\ell)$. Weighing web sites enables users to dynamically adjust the number of web sites over which they can be spatially and/or temporally tracked depending on the cumulative privacy

[3] n-grams are consecutive character sequences of length n.

cost. In addition, advertisers can spatially track users across *different categories*.

The TP-cookie management policies specify that a TP-cookie may be used with a number of associated web sites as long as its privacy cost is limited:

- **Spatial tracking policy:** For each new web site $b_\ell \notin H(B)$ associated with a known hidden server $d_j \in H(D)$, the weights $\omega_1(b_\ell)$ and $\omega_2(b_\ell)$ are automatically determined. An existing TP-cookie $c_k \in H(K)$ assigned by d_j can be associated with a request to the new web site b_ℓ if the following condition holds:

$$\sum_{b_m \in H(B)} \gamma(b_m, d_j, c_k) < L_S \tag{3}$$

  where L_S is the maximum privacy cost allowed for a cookie. Otherwise, c_k is not sent and a new TP-cookie will be assigned by the third-party d_j.
- **Temporal tracking policy:** For a known web site $b_\ell \in H(B)$ connected to a third-party server $d_j \in H(D)$, the same TP-cookie $c_k \in H(K)$, can be used for the time period L_T or for at most L_V visits:

$$\sum_{b_m \in H(B)} \upsilon(b_m, d_j, c_k) \cdot \omega_2(b_m) < L_V \tag{4}$$

  The time period during which a user can be profiled now depends on the weight of the web site.

Consider the following example. User u visits four web sites sequentially: b_1: *www.google.com*, b_2: *www.google.ch/search?q=computers*, b_3: *www.facebook.com*, and b_4:*www.facebook.coms.php?q=nevena&sid=2a1f75*. We assume that all web sites are connected to the same third-party d_j, that there are no TP-cookies in the browser initially, and that $N_C = 10$, $L_S = 0.5$, and $L_V = 5$.

The weights $(\omega_1(b_\ell), \omega_2(b_\ell))$ are automatically computed: (3, 0.1), (3, 0.9), (10, 0.1), (10, 1) for b_1 to b_4 respectively. The weights $\omega_1(b_\ell)$ depend on user preferences. In this example, we observe that user u is unwilling to reveal information about his social networks and assigns a high weight to $facebook.com$. The weights $\omega_2(b_\ell)$ are computed based on the URLs. For generic URLs (b_1 and b_3), the weights are low, whereas for specific URLs (b_2 and b_4), the weights are high because they contain keywords that reflect the user's interests.

The policy for spatial tracking allows the same TP-cookie to be sent for the three web sites b_1, b_2 and b_3 as their cumulative privacy cost is below the threshold: $\sum_{m=1}^{3} \gamma(b_m, d_j, c_k) = (0.1 \cdot 3 + 0.9 \cdot 3 + 0.1 \cdot 10)/10 = 0.4 < 0.5$. However, the fourth web site b_4 requires a separate TP-cookie for this URL as its privacy cost is too high: $\gamma(b_4, d_j, c_k) = 1$. Note that the spatial policy allows a TP-cookie to be shared across 16 web sites of the same category and same URL weight as b_1.

In addition to the spatial policy, the temporal policy must be verified before sharing a TP-cookie across web sites b_1 and b_2. We compute: $\sum_{m=1}^{2} \upsilon(b_m, d_j, c_k) \cdot \omega_2(b_m) = 1 \cdot 0.1 + 1 \cdot 0.9 = 1 < 5$. As it is lower than L_V, the same TP-cookie can be used for b_1 and b_2. Note that the same TP-cookie can be used 50 times with web sites of the same category and same URL weight as b_1, whereas it can be used only 5 times for web sites of the same category and same URL weight as b_2.

Users have a fine-grained control over the dissemination of their personal information and can decide *when*, *where* and *for how long* they will be tracked. Yet, advertisers can track users across categories depending on users' privacy preferences and serve relevant advertisements.

D. Discussion

The third approach is superior to other approaches as it allows for a finer-grained control of the information shared with third parties. To set their preferences on the allowed amount of tracking (L_S, L_T, and L_V), and on web site categories, users have two possibilities: (i) Users can *manually* set their preferences for each parameter and each category [12], or (ii) users can *automatically* define their preferences supported by online social communities. In particular, users can reuse profiles of preferences created by other users. Recently, there have been several efforts to support the privacy management via community expertise [1], [5], [23]. For example, Ad Block Plus [1] (an advertisements blocking extension) is based on a subscription service: Users of the extension can automatically download lists of URLs to block from other users.

IV. Implementation

In order to test our approach, we implemented an extension of the Firefox web browser called *PrivaCookie* as a proof of concept.[4] In this section, we first explain the main challenges of the implementation, describe our study, and provide results.

A. Cookie Management

The extension first detects cookies sent to third-parties and then applies the privacy-friendly cookie management proposed in the previous section.

1) Third-Party Cookies Detection: Our extension detects cookies sent to third-parties by comparing the URL of the current HTTP connection with the URL of the server that caused the connection. To do this, Firefox provides objects and interfaces, namely *nsIChannel* and *nsICookiePermission*. Starting with Firefox 3, the browser remembers with the function *getOriginatingURI* of *nsICookiePermission* the *originating server* of each connection (i.e., the server that caused the connection). Hence, TP-cookies are detected by analyzing every outgoing connection to a server (*nsIChannel*) and determining whether the destination corresponds to the originating server or to a third-party server. In other words, by simply comparing the URL of the current connection with the originating URL, we determine whether the connection is directed toward a third-party server, and implicitly determine whether cookies sent over this connection are TP-cookies. This method is used by Firefox to properly detect TP-cookies. With this method, cookies sent to a first-party server in the past, and then sent to third-party servers are also detected. In our extension, all detected TP-cookies are stored in a local table. Note that, in the current implementation, we do not parse packets to find

[4]The code is available at http://icapeople.epfl.ch/freudiger .

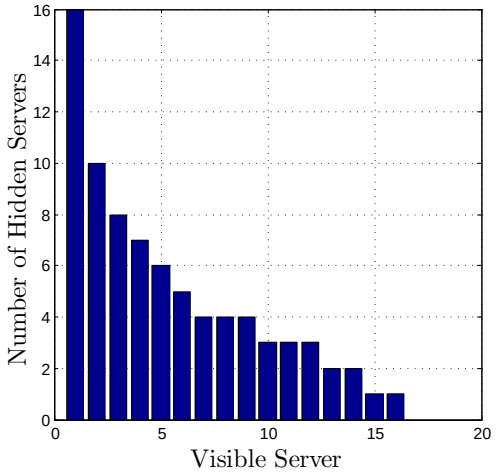

Fig. 3. Number of hidden servers for each of the top 20 web domains.

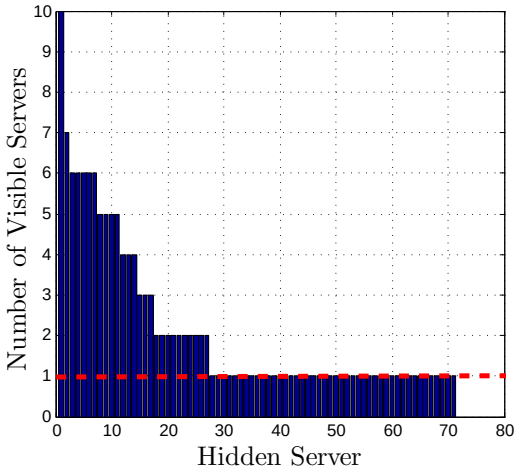

Fig. 4. Number of visible servers for each hidden server. The dashed line represents the limitation of the number of associations imposed by the extension.

TP-cookies set in Javascript. TP-cookies set in Javascript are thus simply blocked. However, as shown in the results, TP-cookies set in Javascript represent a minority of TP-cookies.

2) Cookie Management Strategy: The extension implements the *no spatial tracking across domains and limited temporal tracking* policy. It first intercepts all TP-cookies exiting/entering the system as explained above. If a TP-cookie $c_k \in H(K)$ should not be sent, then the cookie is removed from the exiting HTTP request. The reply from the web server will then contain a new TP-cookie that the extension stores in its local table. Unlike Firefox, the local table remembers for each TP-cookie the corresponding pair of associated web domain and third party (s_i, d_j) (Tab. II). The implementation of the second and third approach will require fetching alternative TP-cookies from the collection of TP-cookies stored in the local table and including them in the exiting HTTP request.

B. Study

In order to gather realistic data about page downloads and obtain a reproducible Internet browsing experience, we use the Firefox browser augmented with the *Pagestats* [18] extension. The extension allows the browser to run in batch mode where a list of sites is specified. We choose 10 pages from each of the top 20 domains across all categories from Alexa's global top sites [3]. A total of 200 web pages was retrieved from a single location in February 2009.[5]

C. Results

First, we investigate the use of TP-cookies online and the amount of tracking by third parties. Then, we evaluate the proposed extension. All the results were gathered with the developed extension.

[5]The data set can also be found at: http://icapeople.epfl.ch/freudiger/.

1) Statistics: According to official specifications [10], the maximum supported size of cookies is 4KB. Hence, cookies can potentially carry a significant amount of information about the state of HTTP connections. However, we observe that the size of TP-cookies is $\sim$ 300B on average. In other words, TP-cookies are mostly used as identifiers and do not carry much information. This allows for the real time manipulation of TP-cookies and does not require extra resources.

Fig. 3 shows the relation between visible servers corresponding to the top 20 domains and hidden servers. We observe that visible servers are connected to a large number of hidden servers: Roughly half of the web domains are affiliated with at least 4 hidden servers. In particular, $AOL.com$ is connected to a total of 16 hidden servers. Out of the 70 hidden servers, we identified a majority of online advertisers (72%). Among the 20 visited web domains, 16 embedded advertisements. Hence, considering the small sample of web sites, the tracking done by third parties is significant.

In Fig. 4, we show the number of connections between visible and hidden servers (and thus the number of associations). By browsing on 200 web pages among the top 20 domains, we contacted 70 different third-party servers. The most popular third-party server is $doubleclick.net$ which is associated with 10 visible servers. Hence, one online advertiser was able to track users on 10 out of the 20 visited domains. Note that 25 hidden servers are at least associated with 2 web domains. Tab. III shows the associations between visible and hidden servers for the most popular domains and advertisers.

We also observed in our study that only 2 out of 20 web domains used redirections to track users online, whereas 16 out of 20 used third-party cookies. Most redirections were actually used to enable users to post content on aggregators (e.g., digg.com) or to embed third-party content on web pages (e.g., youtube videos). This confirms that tracking based on TP-cookies is the primary concern for the privacy of users with respect to online advertisers.

Hidden Server	Visible Server									
	Yahoo	Ebay	AOL	IMDB	Orkut	Msn	Myspace	HI5	Blogspot	Rapidshare
doubleclick.net	$c_1 \vert c_{1,1}$		$c_1 \vert c_{1,2}$	$c_1 \vert c_{1,3}$	$c_1 \vert c_{1,4}$	$c_1 \vert c_{1,5}$	$c_1 \vert c_{1,6}$	$c_1 \vert c_{1,7}$	$c_1 \vert c_{1,8}$	
quantaserve.net			$c_2 \vert c_{2,1}$			$c_2 \vert c_{2,2}$	$c_2 \vert c_{2,3}$		$c_2 \vert c_{2,4}$	
atmdt.com	$c_3 \vert c_{3,1}$	$c_3 \vert c_{3,2}$	$c_3 \vert c_{3,3}$	$c_3 \vert c_{3,4}$		$c_3 \vert c_{3,5}$				
advertising.com			$c_4 \vert c_{4,1}$	$c_4 \vert c_{4,2}$						
yieldmanager.com	$c_5 \vert c_{5,1}$		$c_5 \vert c_{5,2}$		$c_5 \vert c_{5,3}$		$c_5 \vert c_{5,4}$	$c_5 \vert c_{5,5}$		$c_5 \vert c_{5,6}$

TABLE III

TOP 10 ASSOCIATED VISIBLE SERVERS CONNECTED WITH THE MOST POPULAR ADVERTISERS. $c_1 \vert c_{1,i}$ REFERS TO THE TP-COOKIES ASSIGNED WITHOUT|WITH THE EXTENSION FOR EACH VISIBLE SERVER i.

Finally, we evaluate the number of TP-cookies set in Javascript by comparing the local table of TP-cookies of the extension with the table of cookies of Firefox. We obtain that only 4% of the TP-cookies are set in Javascript, thus indicating that our current approach captures the majority of TP-cookies.

2) Success of the Extension: The extension generates and maintains a collection of TP-cookies for each third-party server based on the pair (s_i, d_j). The current implementation does not allow for spatial tracking and thus the size of associations is limited to 1 (dashed line in Fig. 4). In other words, each TP-cookie can be used with only one pair (s_i, d_j). For example in Tab. III, the TP-cookie c_1 of $doubleclick.net$ is replaced with a new TP-cookie $c_{1,i}$ for each associated visible server i. Hence, the size of the collection of TP-cookies depends on the number of associations. In this study, the extension caused 81 additional TP-cookies assignments. Compared with the "block all" solution, our extension in its current form allows for tracking by a single advertiser on a single domain over a limited period of time. The extension works as expected and demonstrates the feasibility of our approach.

V. ONLINE ADVERTISERS COUNTERMEASURES

Online advertisers might consider other tracking techniques to circumvent the privacy-friendly cookie management proposed in this paper.

Online advertisers can track users by their IP address. However, there are various drawbacks with this tracking technique. First, web servers must remember the IP address of each connection, i.e., the state of their connection, in contradiction with the current design of the Internet [10]. Second, an IP address may not only identify a single computer, but could also refer to a computer network using Network Address Translation (NAT).[6] Third, because there are not enough IP addresses to cover the number of users, many ISPs have resorted to the use of dynamic IP addresses. This means that users could be assigned a different IP address every time they access the Internet. Finally, IP traffic can be anonymized using either Tor [19] or an anonymizer [4]. In other words, IP addresses may be unreliable to track users online.

The cache of the web browser also permits to track users online. Juels *et al.* [27] propose to use the cache to store persistent, server-accessible data object called *cache-cookies*. Jackson *et al.* [25] show that because the access to cached elements is not restricted, a web server can verify the presence in the cache of an object from another web site and thus track users across different web sites. To prevent cache tracking, Jackson suggests to regulate the access to the cache by implementing the *same-origin principle* for cache cookies: Only the server that puts a file in a browser cache can access it later. However, this does not avoid tracking by third parties.

The browser history can also be used to track users by exploiting visited URLs stored in the browser [35]. Jackson *et al.* [25] show that the access to the browser history can be regulated by the same-origin policy. Jakobsson *et al.* [26] makes use of the browser history property as a feature for privacy-friendly tracking. Web servers aggregate information from users' history in a privacy-friendly manner. However, users must trust that web servers will not abuse the system.

Plugins (particularly Flash) are another obstacle to online privacy because their own policies may be more permissive than those of web browsers. For example, plugins make use of their own cookies not managed by web browsers. Hence, general policies of browsers do not apply.

The privacy-friendly TP-cookie management proposed in this paper can be applied to solve these problems, thus letting users control the amount of shared information with third-parties.

VI. RELATED WORK

The use of cookies is regulated in Europe [22] and the USA [17]. These regulations define strict rules on the collection, setting and use of cookies. For example, storing cookies in a user's computer is allowed only if: (i) The user is provided information about how this data is used; and (ii) the user is given the possibility of denying this storing operation. However, these regulations are insufficient to protect the privacy of users online as they mostly focus on clarifying the use of cookies.

Shankar and Karlof [34] propose Doppelganger, a Firefox extension to manage cookies. Users only have to make a small number of high-level decisions to manage their cookies. The value of a cookie is determined visually by comparing a web page with and without the FP-cookie. TP-cookies are, however, systematically blocked.

[6]NAT enables multiple hosts on a private network to access the Internet using a single IP address.

Krishnamurthy and Malandrino [28] propose to filter the data exiting web browsers. They suggest a binary management of the information: Block or allow. They investigate the trade-off between web pages usability and privacy and show that blocking third-party cookies reduces tracking online without affecting the usability of web pages. Hence, disabling third-party cookies is a good solution for the privacy of users. However, it entirely impedes behavioral advertising, making advertisements less relevant to the user interest [32]. Hence, we consider an approach that studies the trade-off between advertising customization and privacy.

The same-origin principle is another spatially restrictive policy used by other extensions [25]. However, it is too permissive to prevent third-party tracking. It allows a TP-cookie to be sent to a third-party for an unlimited number of associated web sites. Our solution complements the same-origin principle by limiting the re-use of TP-cookies.

The support of privacy management by a social community is suggested by Goecks and Mynatt [23]. The authors develop a tool called Acumen that users can consult to improve their privacy decisions. We rely on similar mechanisms for the definition of user preferences.

Recently, online advertisers developed tools that lets users choose interest categories to improve the relevance of advertising [12]. With this approach, besides observing browsing activities, online advertisers also get additional personal information. Instead, with our solution, users can still choose interest categories to obtain relevant advertisements, while sharing less information with advertisers.

VII. Conclusions and Future Work

We have considered the trade-off between advertising customization and online tracking of users. We have proposed a novel approach to handle TP-cookies that enables users to control the amount of information shared with advertisers. To do so, our solution maintains a collection of TP-cookies for each advertiser. The decision to use a given TP-cookie is based on a cost-benefit analysis that depends on the visited web site and the value of the TP-cookie. To valuate TP-cookies, we considered three approaches that take into account user privacy preferences and differ in the achieved trade-offs. We have evaluated the feasibility of our solution by implementing a Firefox extension. Our solution empowers users to manage TP-cookies in a privacy-friendly manner. Hence, together with Doppelganger, our extension provides a complete privacy-friendly management of cookies.

We plan to implement the advanced cookie management approaches and improve the handling of TP-cookies set in Javascript. We also intend to consider other criteria, such as users' level of trust in different advertisers.

Acknowledgments

We would like to thank Maxim Raya, Marcin Poturalski, Mark Felegyhazi, Reza Shokri, and the anonymous reviewers for their helpful feedback on earlier versions of this work. Special thanks go to Fabien Dutoit and Aurélia Rochat who implemented the extension.

References

[1] Adblock plus. http://www.adblockplus.org.
[2] Adtech optout cookie. http://www.adtech.info/cookie_opt-out/.
[3] Alexa: Most popular web sites. http://www.alexa.com.
[4] Anonymizer: Online privacy and security. http://www.anonymizer.com.
[5] Cookiepedia. http://cookies.softwareblaze.com/Cookiepedia.
[6] Noscript. https://addons.mozilla.org/firefox/722/.
[7] Privoxy. https://www.privoxy.org.
[8] Top 25 website categories by advertising revenue in 2006. http://blog.econsultant.com/top-25-website-categories-by-advertising-revenue-2006-tns-media-intelligence.
[9] Url categories. http://www.websense.com/content/URLCategories.aspx.
[10] RFC 2109. Http state management mechanism. http://www.ietf.org/rfc/rfc2109.txt.
[11] E. Baykan, M. Henzinger, and I. Weber. Web page language identification based on urls. In *VLDB*, 2008.
[12] The Official Google Blog. Making ads more interesting, March 2009.
[13] D. Cancel. Ghostery watches the web sites that are watching you. https://addons.mozilla.org/en-US/firefox/addon/9609, March 2009.
[14] M. Chew, D. Balfanz, and B. Laurie. (under)mining privacy in social networks. In *Web 2.0 Security and Privacy*, 2008.
[15] R. Chow, P. Golle, and J. Staddon. Inference detection technology for web 2.0. In *Web 2.0 Security and Privacy*, 2007.
[16] M. Christodorescu. Private use of untrusted web servers via opportunistic encryption. In *Web 2.0 Security and Privacy*, 2008.
[17] Federal Trade Commission. Online behavioral advertising: Moving the discussion forward to possible self-regulartory principles. 2008.
[18] S. DeDeo. Pagestats. http://www.cs.wpi.edu/~cew/pagestats/, 2006.
[19] R. Dingledine, N. Mathewson, and P. Syverson. Tor: the second-generation onion router. In *USENIX Security Symposium*, pages 21–21, 2004.
[20] eMarketer.com. Behavioral advertising on target... to explode online. http://www.emarketer.com/Article.aspx?id=1004989, June 2007.
[21] eMarketer.com. Which online ads get attention. http://www.emarketer.com/Article.aspx?id=1007003, March 2009.
[22] EU. On the protection of individuals with regard to the processing of personal data and on the free movement of such data. EU Data Protection Directive 95/46/EC, October 1995.
[23] J. Goecks and E. D. Mynatt. Supporting privacy management via community experience and expertise. In *Conference on Communities and Technology*, 2005.
[24] IAB. Behavioral targeting: Secret weapon in display ad's arsenal. July 2008.
[25] C. Jackson, A. Bortz, D. Boneh, and J. C. Mitchell. Protecting browser state from web privacy attacks. In *WWW*, 2006.
[26] M. Jakobsson, A. Juels, and J. Ratkiewicz. Privacy preserving history mining for web browsers. In *Web 2.0 Security and Privacy*, 2008.
[27] A. Juels, M. Jakobsson, and T. N. Jagatic. Cache cookies for browser authentication. In *IEEE Symposium on Security and Privacy*, 2006.
[28] B. Krishnamurthy, D. Malandrino, and C. E. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. In *Symposium on Usable privacy and security*, 2007.
[29] B. Krishnamurthy and C. E. Wills. Generating a privacy footprint on the internet. In *IMC*, 2006.
[30] B. Krishnamurthy and C. E. Wills. Characterizing privacy in online social networks. In *SIGCOMM Workshop on Online Social Networks*, 2008.
[31] PricewaterhouseCoopers. IAB internet advertising revenue report. March 2009.
[32] RevenueScience. Sixty three percent of consumers always prefer advertising based on their interests. Press Releases, April 2006.
[33] R. T. Rust and S. Varki. Rising from the ashes of advertising. *Journal of Business Research*, 37:173–181, 1996.
[34] U. Shankar and C. Karlof. Doppelganer: Better browser privacy without the bother. In *CCS*, 2006.
[35] L. Weinstein. New web analytics service spies on web browsing activity without permission. http://lauren.vortex.com/archive/000498.html.