Privacy-Preserving and Advertising-Friendly Web Surfing

David Sánchez and Alexandre Viejo[1]

Universitat Rovira i Virgili, Department of Computer Science and Mathematics, UNESCO Chair in Data Privacy, CYBERCAT-Center for Cybersecurity Research of Catalonia, Avda. Països Catalans, 26, 43007 Tarragona, Spain

Abstract

Behavioral targeting is a technique extensively used by the advertising industry that consists on compiling the interests of the individuals that surf the web by tracking the web sites they visit. This targeting strategy has proven to be, by far, the most effective form of advertising and, nowadays, constitutes the pillar of the Internet business model, in which free content is offered to the users in exchange for showing advertisements. Nevertheless, threatening the privacy of the people by tracking them and gathering their interests and habits is not a minor issue and, in recent years, many users have reacted by installing tools in their browsers that block all connections suspicious of being advertisements. By blocking indiscriminately and systematically, these tools protect the privacy of their users, but also hamper the online advertising business and, hence, endanger the sustainability of the 'free' Internet model. To tackle this issue, we propose a system that conciliates users' privacy during web surfing and the advertising business. Specifically, our proposal empowers the user in the protection of her privacy by allowing her to define her privacy requirements, that is, which user's interests should be hidden from the advertising platforms and which ones can be revealed. Then, our system selectively blocks or bypasses tracking on the browsed web sites according to their content and the privacy requirements. Our system is also the first that implements simulated browsing sessions as a privacy-preserving measure. Empirical results show that our proposal allows fine grained and user tailored privacy protection, which is also more accurate, responsive and respectful with the Internet business model than related works.

Keywords: Privacy, Web surfing, Web tracking, User profiling, Online advertising.

1. Introduction

Since 1995 the penetration of the Internet in the Society has grown dramatically and it has consolidated itself as a main platform of services that has changed the way people communicate, do business or enjoy their leisure time. The entry point for any Internet user is the so-called *web surfing*; this is, navigate through the World Wide Web by means of a web browser, retrieving and consuming web pages filled with information, images, videos and other sorts of content.

Noticing the business opportunity, the advertising industry has deployed important efforts in the digital arena to gather economic benefits from the huge quantity of individuals surfing the web. As a matter of fact, Internet advertising revenues in the US increased a 22% in 2016 from a year earlier to a record of $72.5 billion, surpassing for the first time in history the amount spent on TV ads [1].

As the revenues of the online advertising industry have grown greatly, their tools for targeting advertisements to the right individuals have also improved significantly. Simple targeting approaches are

[1] Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona. Spain
Tel.: +034 977 558270; Fax: +034 977 559710;
E-mail: alexandre.viejo@urv.cat

based on showing advertisements to persons according to their location or to the contents of the web pages they browse. More sophisticated tools, such as the well-known *behavioral targeting*, rely on compiling the interests of the individuals, which have proved to be, by far, the most effective form of advertising [2]. In particular, behavioral advertising has allowed the advertising industry to remarkably increase their rates resulting from converting web visitors into paying customers.

Nevertheless, the significant benefits of behavioral advertising do not come without cost. As the literature on this topic has recently revealed, behavioral targeting clearly jeopardizes the privacy of the individuals surfing the web, due to the fact that this strategy is based on tracking and profiling the browsing habits of the people, generally without their knowledge and consent [3]. Nowadays, there is a social understanding of the privacy threats that the uncontrolled collection and exploitation of these personal data may produce, which include discriminatory actions or unethical exploitation of data [4]. In this respect, the recent E.U. GDPR[2] has put the spotlight on the privacy risks that such activities pose to the Society, and has requested the implementation of technological solutions that not only protect the privacy of the individuals, but also improve transparency and empower the users on the management and protection of their data.

Current privacy-preserving tools, such as Adblock Plus[3] and Ghostery[4], have been proved to be effective in blocking advertisements and avoiding tracking. Nevertheless, these solutions lack from flexibility because they just block any content that may be suspicious of being an advertisement or allowing any sort of tracking. While this is fine from the privacy point of view, it also represents a serious harm to the dominant advertising revenue model, which is based on offering free Internet content and services to the individuals in exchange of showing advertisements to them [5, 6]. The industry states that, currently, a 90% of the web sites are free only because they use advertisements to sustain their business model [7]. In this way, a generalization of the use of this 'systematic' blocking tools would lead companies to apply paid alternatives such as the *Freemium model*, the *Fee-for-service model* or the *Subscription model*, all of them based on offering paid services and content [8]. This potential shift from a 'free' Internet to a paid one has raised social concerns, because it potentially endangers the sustainability of the current model that maintains the Internet open to everyone [6, 9]. More specifically, recent surveys [10] have shown that, in December of 2016, a 11% of the global internet population was blocking ads on the web, which is over 600 million devices running ad block software globally. Even more relevant is the fact that in, one year, these studies have reported a 30% of global grow in ad block usage. The fast penetration of ad blocking in the society has already brought a negative effect on the publishers of online free content that have reported billions of advertising revenue lost.

As a first sight of the emergent war between ad blockers and the advertising industry, the literature reports that the arrival of ad blocking tools has already forced online advertising companies to adopt new strategies; specifically, several of these entities have started to apply countermeasures that allow their advertisements to be unblocked, or even they have started to deny access into their controlled web sites to the users of these tools [11, 12], an action that goes against the spirit of the forthcoming E.U. ePrivacy Regulation[5]. Another significant movement aligned with this change of strategy is the decision of Google, one of the biggest companies in the advertising business, of providing its well-known web browser with an ad blocker natively integrated. As reported in the literature, this decision may allow Google to use its power in the browser market (Google*'s Chrome* has been reported to own a 59.2% of the web browser market share[6] in November'17) to give itself a tool to decide which advertisements should be blocked and which ones should be allowed [13].

Last but not least, a secondary and generally overlooked side-effect of the use of systematic blocking tools is the fact that they also block advertisements that may be useful for the users. As it was brought into attention in [14], advertisements that match users' profiles may be useful and legit sources of information that warn users about new products and trends of their interest; the use of rigid blocking tools clearly impedes this usage and its related benefits.

According to these points, and in order to balance the people's right to the privacy, the economic benefits of the companies that support the Internet, and the use of advertisements as a useful source of information, current legal frameworks and the Society in general demand new privacy-preserving tools that empower individuals -allowing them to regain control over their privacy while surfing the Internet-, and that behave in a respectful way with the current Internet business model -paving the way for an Internet still open and free for everyone-.

---

[2] E.U. General Data Protection Regulation. https://www.eugdpr.org/ (last accessed: 09/07/2018)

[3] Adblock plus. https://adblockplus.org (last accessed 09/07/2018)

[4] Ghostery. https://www.ghostery.com (last accessed 09/07/2018)

[5] E.U. ePrivacy Regulation. https://ec.europa.eu/digital-single-market/en/proposal-eprivacy-regulation (last accessed: 09/07/2018)

[6] W3Counter. https://www.w3counter.com/globalstats.php (last accessed: 09/07/2018)

1.1. Contribution and plan of this paper

In this paper, we propose a system that conciliates users' privacy during web surfing in front of behavioral targeting and the ad-based business model that sustains the Internet.

Our proposal empowers the user in the protection of her privacy by allowing her defining her privacy requirements with respect to advertising platforms; these requirements specify which personal interests may be revelated to the advertising platforms (and, therefore, be used to conduct targeted advertising) and which interests should remain hidden.

Then, our system, which is meant to be deployed as a web browser plugin, constantly monitors the user's web browsing requests and the tracking implemented by the advertising platforms, to enforce privacy-preserving measures consistent with the privacy requirements. In contrast with tools that systematically block all tracking, our system selectively blocks or bypasses tracking on the browsed web sites according to their content (which may reveal user's interests) and the privacy requirements of the user. In addition to blocking tracking, our system is also the first that implements simulated (tracked) browsing sessions as a privacy-preserving measure. With this, we aim at limiting the amount of blocking needed to preserve the user's privacy, thus reducing the harm in the ad-based business.

Empirical results conducted on a variety of scenarios (i.e., privacy requirements) show that our system allows fine grained and user tailored privacy protection, which is also more accurate, responsive and respectful with the Internet business model than related works.

The rest of the paper is organized as follows. Section 2 details the main entities and notions involved in the online advertising model. Section 3 presents our system and details its components and algorithms. Section 4 reports and discusses the results of the empirical experiments, and compares them against several baselines. Section 5 reviews related works on privacy protection in online advertising. The final section enounces the conclusions of this work.

2. The online advertising model

The online advertising industry has received significant attention in recent years. Some authors have analyzed this environment and identified the key aspects that support the online advertising model. In particular, a very deep analysis can be found in [15], which has served as a basis to other works such as [3] and [14]. In this section, we follow the same line, and briefly introduce the key elements that allow understanding the insights of our proposal.

As explained in the introduction, advertisers may implement different targeting approaches to decide which advertisements are best suited to web surfers. In plain words, the 'best' advertisement is the one that has better chances of converting a web surfer into a paying costumer. The process of selecting that advertisement is executed under the terms of the targeting approach in use. The most usual targeting strategies are: *location-based*, *contextual-based* and *profile-based*. The first strategy is based on showing advertisements that suit with the person's location. The second strategy shows advertisements that are related to the content of the visited web site. Finally, the last approach, also known as *behavioral targeting,* gathers the interests of the web surfers, uses this information to build so-called *user profiles*, and exploits them to select and show advertisements related to the dominant interests of web users.

These targeting strategies, together with the classic *generic approach*, in which advertisements are shown to the web surfer without applying any sort of data processing, are the tools of the *online advertising model*; this model is composed of three main actors and of the advertisement-serving process that, finally, delivers the selected advertisement to the selected person.

The three actors are next detailed:

- *Publisher*: it is the entity that owns a web site and provides placement for advertisements in exchange for economic benefits.
- *Advertiser*: it is the entity that is willing to show advertisements on the placements offered by the publishers according to the targeting criteria best suited for its marketing campaign, and in exchange for an amount of money.
- *Advertising platform*: it is an entity with a twofold purpose:
    - It is capable of tracking web surfers across the Internet by means of techniques such as Internet cookies, query strings, local shared objects or browser fingerprint [16]. Tracking users allows advertising platforms compiling the so-called users' *clickstreams*, which consist of the sequence of web sites that the users have browsed. Note that each advertising platform performs its own tracking process and, hence, different platforms

may build (slightly) different clickstreams for the same user. Advertising platforms use the compiled clickstreams to extract the data (i.e., location, interests, etc.) required by the targeting approaches explained above. For example, in the case of the profile-based/behavioral targeting, each advertising platform generates its own user profile for each user by extracting her interests from her clickstream [15, 17-21].

- It connects advertisers with publishers, allowing the former to place the best suited advertisement to each user by means of the information (e.g., personal data, profiles) acquired in past steps. It is important to mention that: i) in the online advertising model, several advertising platforms (e.g., Google's DoubleClick[7] or Microsoft's Bing Ads[8]) offer their services to advertisers, and the latter generally buys the services of several platforms simultaneously in order to reach more users; and ii) during the whole process, the advertising platforms are the only entities that have access to the user profiles; hence, only these entities represent a threat to the privacy of the users.

Finally, the advertisement-serving process is a straight-forward procedure in which a user browses a web site hosted by a publisher that contains embedded links to different advertising platforms. The user's browser loads these links, which allow the advertising platforms to run their tracking methods, identify the web surfer and place the proper advertisements according to the objectives of the advertisers and the information of the visitor that the advertising platforms have gathered.

3. Our proposal

This section details the system we propose to conciliate users' privacy during web browsing and the ad-based business model. Our system focuses on protecting the users against the most intrusive and privacy-harmful advertising strategy: *behavioral targeting*, which builds user profiles from the aggregation of evidences on the users' interests extracted from their *clickstreams*.

To protect the users' privacy, that is, to avoid advertising platforms from obtaining accurate characterizations of their personality, our system implements a set of profiling countermeasures. Instead of systematically blocking all tracking, the countermeasures we propose act on the users' clickstreams to introduce a configurable degree of distortion to the profiles built by the advertising platforms that may track them. By configuring the degree of profile distortion, the user can balance the level of privacy protection she is comfortable with (i.e., her privacy requirements) and, also, the utility she gets from the advertisements shown to her (i.e., how accurate the ads are with respect to her real interests). Whereas the former empowers the individuals and allows them to regain control over their privacy while browsing the web, the latter sustains the current advertising business model.

In the following we detail how user profiles are analyzed and managed by our system. Afterwards, we detail the architecture of our system and the algorithm that decides which profile countermeasures are implemented on the user's clickstream to fulfill her privacy requirements.

3.1. Building user profiles

In the area of user profiling [22], profiles are modelled as distributions of topics or *categories* (e.g., sports, science, etc.) that reflect the interest or preference of the user on such topics. When user profiles are built from web browsing, category distributions are obtained by aggregating the evidences on the user's interests gathered from her *clickstream*. The examined evidences may vary but, usually, profilers consider the elements that best and less ambiguously represent the content of the web site that the user decides to browse (such as keywords or titles, as proposed in [14]). These elements, which are assumed to implicitly represent the interests of the user, are classified within the (more general) categories considered in the user profile. The contribution of each element towards its corresponding category is usually quantified as the *term's frequency-inverse document frequency* (TFIDF) [23], which reflects how relevant the element is in the context of the corpus of documents considered by the profiler (e.g., the set of web sites tracked by the advertising platform). In this manner, very general elements that frequently occur (e.g., health) will contribute less to their corresponding profile category than more specific and, thus, less frequent terms (e.g., lung cancer).

[7] DoubleClick. https://www.doubleclickbygoogle.com/ (last accessed: 09/07/2018)
[8] Bing Ads. https://secure.bingads.microsoft.com/ (last accessed: 09/07/2018)

To tackle the scenario that we consider in this paper, we need to estimate the profile characterization that advertising platforms build by analyzing the user's clickstream. In this way, we will be able to implement the privacy-preserving countermeasures required to distort and, therefore, protect the profiles. However, the specificities of the profiling mechanisms implemented by the advertising platforms are usually unknown: concretely, the set of categories considered in the profile or the corpus employed to compute the TFIDF. To be able to cope with different profiling configurations, we employ a profiling mechanism that is more general, i.e., less dependent on the profiling parameters, than the above-described one. In this way, we aim at better estimating the profiles built by the variety of advertising platforms that may track the user and, therefore, at implementing more accurate profiling countermeasures.

Our approach, which is grounded in the Information Theory, quantifies the contribution of the elements extracted from the browsed web sites to their corresponding profile categories according to the semantics they convey. In general, the amount of semantics conveyed by a textual entity can be measured from the amount of information it provides, that is, its *information content* (IC) [24]. To be less dependent on the set of categories considered in the user profile (which may be more or less general, or finer or coarser grained from one platform to another), the contribution of each element belonging to a profile category is weighted by the relative importance, i.e., informativeness, of such category with respect to the others. Moreover, to be less dependent on the corpus used to measure the contribution of each element to the profile category, we employ the largest electronic repository available: the Web. Details on how evidences are gathered from the user's clickstream and on the calculation of their contribution to the estimated profiles built by the platforms are given below.

Formally, let us consider the profile PP_X that is built by the advertising platform X as a tuple of pre-defined categories of interests, $C=\{c_1, ..., c_k\}$, and their relative weights (v_i) in the profile distribution, that is, $PP_X=\{<c_1,v_1>, \ldots, <c_k,v_k>\}$. At the beginning, i.e., the when clickstream is empty, all category weights are initialized to zero: $PP_X^{ini}=\{<c_1,0>, \ldots, <c_k,0>\}$. Whenever the user browses a web site tracked by the platform X, her profile with respect to such platform is updated by executing two steps: i) extraction and classification of evidences from the web site; and ii) updating of category weights.

3.1.1. Extraction and classification of evidences

The web site browsed by the user is parsed to extract relevant evidences k_j on the user's interests from the web content; specifically, *keywords* are gathered, which are highly informative elements that unambiguously describe the content of the web site. If keywords are not available, the noun phrases found within the *titles* of the web site are considered.

The elements k_j are semantically classified within the categories C considered in the profile. Because the categories in C for each platform are unknown, we use the most general categories of a general-purpose knowledge base, which we also use to match the elements to their categories. Specifically, we employ the Open Directory Project (ODP)[9], a collaborative Internet directory that classifies around 5 million web sites within more than 1 million taxonomically organized categories. By querying each extracted element k_j into ODP, we obtain the hierarchy H_j of categories that generalize it; e.g., if we query the element "Windows 7", found as *keyword* in a web site about Microsoft's OS, we obtain the following hierarchy of categories: *Computers* → *Software* → *Operating Systems* → *Windows 7*. Because profile categories correspond to the most general categories in ODP (*arts, business, computers, games, health, home, news, recreation, reference, regional, science, shopping, society, sports, kids & teens, word*), a profile category c_l for k_j is always found within H_j.

3.1.2. Updating of category weights

The contribution of each element k_j to its profile category c_l is quantified by the *amount of information* k_j conveys, that is, $IC(k_j)$. The IC of an element is computed as the inverse logarithm of its probability of occurrence [25].

$$IC(k_j) = -\log p(k_j) \tag{1}$$

Similar to TFIDF approaches, in this way, specific terms that rarely occur are considered more informative than commonly used ones. Nevertheless, on the contrary to profiling mechanisms relying on

[9] Open Directory Project. http://dmoztools.net/ (last accessed: 09/07/2018)

absolute occurrences [26], information theoretic profiles tend to be more general and accurate [27]; in fact, the notion of IC has been widely used as a general metric of the semantics encompassed by linguistic entities in the area of semantic similarity [28].

To be truly general, IC requires robust probabilities computed from corpora representative of the actual distribution of terms as used in Society. The Web, which is the largest and most up-to-date electronic repository currently available, fits this criterion. In fact, the Web is so vast, and the number of authors generating content so enormous (which can be assumed as a representative sample from humankind), that we can assume that the frequency of terms in the Web approximate the actual relative frequencies of those terms in Society [29]. Moreover, probabilities of terms in the Web can be straightforwardly estimated from the *hit count* (i.e., number of results) provided by Web search engines (e.g., Google) when querying the terms of interest [30]. According to these assumptions, we compute the IC of each k_j as follows:

$$IC_{web}(k_j) = -\log \frac{hit\_count(k_j)}{total\_webs} \quad , \tag{2}$$

where *total_webs* stands for the total number of web sites indexed by the Web search engine.

Finally, we compute the contribution of k_j to its profile category c_l as the $IC_{web}(k_j)$ weighted by the informativeness of the category c_l itself, that is, $IC_{web}(c_l)$. Because categories in the user profiles may vary from one advertising platform to another, and categories may have different degrees of generality, this calculation contributes to make the profile characterization more general and independent of the categories C considered by each platform [27]. Formally, the weight v_l of category c_l is updated according to $IC_{web}(k_j)$ as follows:

$$v_l += IC_{web}(k_j) \times IC_{web}(c_l) \tag{3}$$

This compensates the fact that, for general categories, the chance of extracting an evidence belonging to that category is higher than for more specific categories. Therefore, if the profile categories considered by a platform are very general, a larger number of more informative evidences will be required to obtain the same weights as for platforms considering more concrete categories.

3.2. System architecture

Our system consists of a local plugin that runs locally on the user's web browser. It constantly monitors the users' browsing requests and decides which actions should be undertaken on the clickstream in order protect (distort) the profile that the advertising platforms may acquire. To do so, it relies on two main elements:

1. *Privacy requirements*: as stated above, the user may specify her privacy requirements by configuring the degree of distortion to be introduced on the profiles built by the advertising platforms. In this way, the user may balance the trade-off between the protection of her profile and the accuracy of the ads sent by the advertising platforms. To define her privacy requirements, the user is asked to define the *target profile* (TP) she desires to reveal to advertising platforms by stating the relative weight of each profile category. The configured *TP* will be used to guide the enforcement of the profile-protection measures we detail below, which aim at making the profiles built by the platforms tend to *TP*. The frontend of this module have been designed with usability requirements in mind and to be intuitive for end users with no particular technical knowledge. Figure 1 shows a screenshot of the configuration panel (integrated in a *Chrome Extension*), which allows the users to set their *TP* in an intuitive way. Specifically, *TP* is set only once during the initial configuration of the plugin and the user can set the relative weight (v_i) of each category in her target profile by means of continuous sliders in the 0..1 range. The sliders of each category are automatically synchronized so that, after setting any weight, we ensure that $\sum v_i=1$ (see details in Section 3.3). According to the privacy and utility requirements of the user, *TP* can be configured in different ways:
    a. By setting all the weights to identical values (i.e., *balanced TP*), we have a *TP* that does not disclose anything to third parties, because all the categories have the same importance. Obviously, if advertising platforms can only learn this *balanced* profile, the

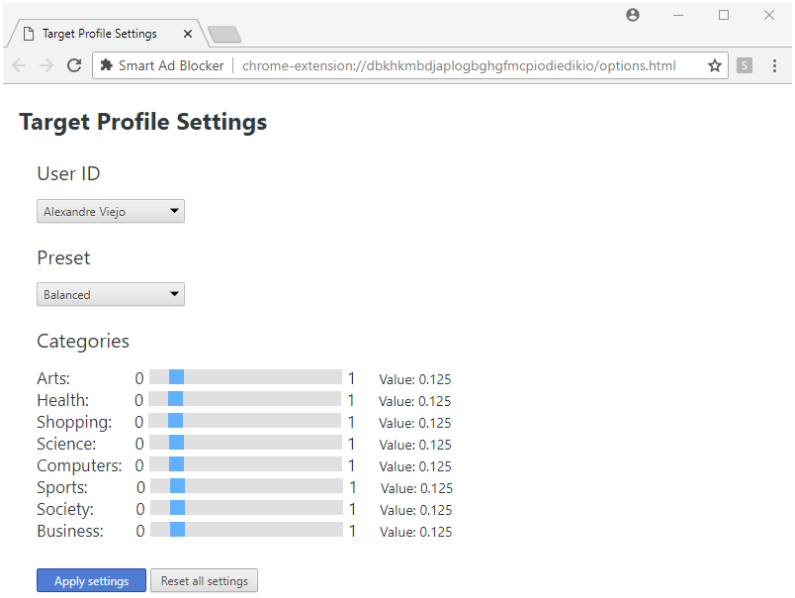
ads they will send to the user will be random (because the user does not seem to have dominant interests), which is something to be consider when setting *TP*.

b. An *unbalanced TP* (i.e., with dominant categories) will enable receiving ads on certain topics. The divergence between the dominant categories *TP* and the real interests of the user (i.e., the dominant interests in her *real profile*, *RP*), will affect to i) the degree of profile protection (i.e., the larger the divergence, the stronger the protection) and ii) the accuracy of the received ads with respect to *RP* (i.e., the larger the divergence, the less interesting the received ads will be for the user).

The frontend also provides presets for the configuration settings above (i.e., a *balanced TP* and several *unbalanced TPs* with dominant categories). If the presets suit the needs of the user, she would not even need to manually set profile weights, thus making the configuration step straightforward.

2. *Estimated profiles built by the advertising platforms:* let us consider that the user's clickstream may be tracked by a set of advertising platforms $\{X_1, \ldots, X_p\}$, e.g., Google's DoubleClick or Microsoft's Bing Ads. PP_{Xi} is the estimated user profile that a specific advertising platform X_i builds from the user's clickstream. Notice that, if no protection is implemented on the clickstream and the platform X_i is able to track *all* the web sites browsed by the user, then $PP_{Xi} = RP$; on the other hand, the goal of our system is to protect the user clickstream so that $PP_{Xi} \rightarrow TP$ for all X_i in $\{X_1, \ldots, X_p\}$.

Figure 1. Configuration panel of our system (integrated in a *Chrome Extension*), which allows users to set their *Target Profiles*. The value of each category represents the relative weight (v_i) of the category in the user profile that is revealed to the advertising platform. Values are normalized so that $\sum v_i = 1$, as detailed in Section 3.3.



To make $PP_{Xi} \rightarrow TP$, our system monitors each browsing request by the user, measures the divergence between the PP_{Xi} it estimates (for all X_i) and *TP,* and implements some of the following actions on the user's clickstream:

- *A1*: to bypass the browsing requested, which means that any tracking implemented in the web site will be effective, and may result in updates on PP_{Xi} for all the platforms tracking the web site. This action will be undertaken when the content of the requested web site does not impair the privacy of the user.
- *A2*: to block the tracking that any or all advertising platforms may have implemented in the requested web site. This implies that the user will be able to browse the web, but no evidences on such browsing will be gathered by the blocked platforms. This action will be undertaken when the content of the web site impairs the privacy of the user with respect to any (or all) advertising platform(s).
- *A3*: to execute a simulated browsing of a randomly chosen web site with specific contents and with non-blocked tracking from concrete platforms. The goal of this action is make the advertising platforms believe that the user has browsed a web site with some specific content (interests) and

produce a controlled update of PP_{Xi} for some (or all) X_i. This action aims at making PP_{Xi} converge faster to TP.

3.3. Profile protection algorithm

Each time the user requests a web site, a greedy algorithm analyzes the request and the estimated PP_{Xi} of the platforms tracking the web site, and undertakes any of the possible actions detailed above to make $PP_{Xi} \rightarrow TP$. The steps of the algorithm are detailed below.

Let us assume that the user has configured TP according to her privacy and utility requirements and that PP_{Xi} has been initialized to $PP_{Xi}^{ini}=\{<c_1,0>, \ldots, <c_k,0>\}$ for all the advertising platforms $\{X_1, \ldots, X_p\}$ that may track the user. Our system updates PP_{Xi} for each (real or simulated) non-blocked browsing of web sites that include tracking from X_i, by following the procedure detailed in Section 3.1.

Let us assume that the user requests a web site W, which contains a set of evidences on her interests (e.g., keywords), $K=\{k_1,...k_n\}$, and that W is tracked by some of (or all) the advertising platforms in $\{X_1,..., X_p\}$. Then, for each platform X_i tracking W, the system executes the following steps:

1) It measures the divergence between PP_{Xi} and TP. To properly compare category weights, these are normalized according to their relative dominance in the profile. This is done by dividing each category weight v_i in PP_{Xi} and TP by the sum of weights of all the categories, so that $\Sigma v_i=1$. Then, the system measures the difference between each category weight v_i in PP_X and TP, that is, $\Delta_{ci} = (v_i; PP_X) - (v_i; TP)$. According to the sign of the difference, we distinguish three scenarios:
    a. If Δ_{ci} is negative, it indicates that the category c_i is underweighted in PP_X. To compensate this negative divergence, additional web sites tracked by X_i with contents related to c_i should be browsed by the user.
    b. If Δ_{ci} is positive, it indicates that the category c_i is over-weighted in PP_X. To compensate this positive divergence, tracking of further browsing on web sites with contents related to c_i should be blocked for X_i.
    c. If Δ_{ci} is zero, then the category matches the tracked and target profiles and no actions on the browsing of web sites related to c_i are needed.
2) Once Δ_{ci} is computed for all c_i, the system selects the category for which Δ_{ci} is maximum, that is, $c_{max}=argmax\ \Delta_{ci}$ for all c_i in C. The idea is to center the action to be undertaken by the system on the category for which the divergence between PP_X and TP is the largest, because it is the one that requires more "effort" to make it converge toward TP. As detailed in *step 1)*, according to the sign of Δ_{cmax}, three actions are performed on the user clickstream:
    a. If Δ_{cmax} is negative, then
        i. if any of the evidences $\{k_1,...k_n\}$ in the web site W belongs to c_{max} (which means that the next web browsing by the user would contribute to make c_{max} converge toward TP), the system undertakes action *A1*: it does not block the tracking of platform X_i on W. The evidences in W are analyzed as detailed in Section 3.1 and PP_{Xi} is updated accordingly.
        ii. otherwise, the system undertakes action *A3*: it executes a simulated browsing of a new web site W' tracked by X_i whose content matches c_{max}. To do so, it randomly chooses a web site categorized under c_{max} in ODP that implements tracking by X_i, and blocks any tracking by other platforms on W'. This action contributes to make c_{max} converge significantly faster than waiting for the user to access to a web site matching c_{max}. The evidences in W' are analyzed as detailed in Section 3.1 and PP_{Xi} is updated accordingly. After that, the system goes back to *step 1)* to re-evaluate W (i.e., the actual requested web) on the updated PP_{Xi}, which may result in new actions. To avoid an excessive number of simulated browsing (which may cause undesirable bandwidth overhead), the user may set a maximum ratio r between the number of simulated and real browsing (e.g., 1:1, 2:1, or even 0:1). In such case, if the system is in *step 3.a.ii)* and the ratio has been consumed, it tries to balance the next most divergent category (which may not require simulated accesses) by going back to *step 3)* with $c'_{max}= argmax\ \Delta_{ci}$ for all c_i in C-c_{max}. If the same outcome happens for all the remaining categories and for all the advertising platforms tracking W, then W is left unblocked for all platforms.
    b. If Δ_{cmax} is positive, then

i. if any of the evidences $\{k_1,...k_n\}$ in W belongs to c_{max}, which means that the next web browsing may increase the divergence of c_{max}, the system undertakes action *A2*: it blocks the tracking of platform X_i in W.

ii. otherwise, no specific action is undertaken for c_{max} and the system tries to balance the next most divergent category by going back to *step 3)* with c'_{max}= *argmax* Δ_{ci} for all c_i in C-c_{max}. If the same outcome happens for all the remaining categories and for all the advertising platforms tracking W, then W is left unblocked for all platforms.

c. If Δ_{cmax} is zero, it means that *all* the profile categories have the same weight in PP_{Xi} and *TP*. Therefore, no countermeasures are implemented and the browsing on W is bypassed without blocking tracking by X_i (action *A1*). This action, in turn, would likely cause some divergences between PP_{Xi} and *TP*, which will be corrected in further iterations.

After the execution of the two steps above on all the advertising platforms tracking W, the web site W is presented to the user (who will browse it), but some of (or all) the platforms tracking W may have been blocked by the system. Moreover, a number up to r simulated browsing sessions on randomly chosen web sites W' may have been executed to make some negatively diverging categories converge toward *TP*. Note that simulated browsing should not be systematically executed before or after accessing W, since this may produce a deterministic browsing pattern that could be easily detected by advertising platforms. Instead, a random pattern or a pattern that tries to mimic human browsing behaviors should be implemented. Note that this concern has already been considered in works such as [31] and [32].

4. Evaluation

In this section, we evaluate the capability of our system of protecting user profiles according to a variety of privacy requirements, that is, in making PP_{Xi} approximate *TP*. We next detail how the evaluation test bed has been configured:

- As done in related works on user profiling [27, 32], we used as profile categories *C,* eight well-differentiated root categories from ODP: *Arts*, *Health*, *Shopping*, *Science*, *Computers*, *Sports*, *Society* and *Business*.
- We generated the clickstream of a synthetic user as follows. First, we took 100 random web queries from the AOL data set[10], which contains the query logs that thousands of real users submitted during 3 months to the AOL search engine. Our assumption is that web queries represent the interests of the users requesting them. The 100 queries were randomly taken from the whole AOL data set, so that they reflect the aggregated distribution of topics of interest for *all* the users in the data set. Then, we queried the 100 queries into Google's search engine and retrieved the first web sites in the results pages. With the set of 100 web sites we obtained we built the clickstream, $\{W_1,\ldots,W_m\}$, that we use in the experiments depicted below. Notice that according to [32], 100 evidences are enough to build representative and stable user profiles for standard users. The second column in Table 1 depicts the normalized distribution of the profile obtained from the clickstream (that is, the real user profile *RP*) by following the procedure detailed in Section 3.1.
- We configured 3 scenarios that represent well-differentiated privacy/utility requirements by setting *TP* with respect to the *RP* built from the synthetic clickstream:
    - *Balanced TP*: category weights in *TP* have been set to identical values, as shown in the third column of Table 1. As introduced in Section 3.2., in this way, *TP* does not disclose anything to third parties (i.e., maximum privacy) because the profile they are able to learn does not have dominant interests.
    - *Opposite TP*: category weights in *TP* have been set by iteratively swapping the two most distant unswapped weights in *RP*, as shown in the fourth column of Table 1. In this way, *TP* significantly differs from *RP* and, in fact, it will make the user appear to have interests opposite to her real ones (i.e., high privacy).
    - *Similar TP*: category weights in *TP* have been set by iteratively swapping the two most similar unswapped weights in *RP*, as shown in fifth column of Table 1. In this way, *TP* will make the user appear to have interests similar to the real ones (i.e., low privacy). The

[10] AOL Search Data: https://archive.org/details/AOL_search_data_leak_2006 (last accessed: 09/07/2018)

advantage of this scenario is that, on the contrary to the former ones, the ads that the user may receive should still match her interests.

- For evaluation purposes, we considered a worst-case scenario in which a single advertising platform X_i is able to track the whole clickstream of the user. In this scenario, if no countermeasures are implemented, PP_{Xi} will tend to *RP*; that is, the platform will accurately learn the real interests of the user. On the other hand, our system will need to enforce the greatest number of countermeasures to make PP_{Xi} converge to *TP* (especially when *TP* significantly diverges from *RP*).

Table 1. Normalized distribution of category weights in the *RP* of a synthetic user (according to a clickstream of 100 web sites created from AOL queries) and three configurations of *TP*: *balanced*, *opposite* and *similar*.

Profile Category	*RP normalized weights*	*Balanced TP normalized weights*	*Opposite TP normalized weights*	*Balanced TP normalized weights*
Arts	0.227	0.125	0.021	0.22
Health	0.119	0.125	0.16	0.045
Shopping	0.16	0.125	0.119	0.178
Science	0.21	0.125	0.227	0.03
Computers	0.22	0.125	0.03	0.227
Sports	0.3	0.125	0.22	0.021
Society	0.45	0.125	0.178	0.119
Business	0.178	0.125	0.045	0.16

The experiments we conducted for the three scenarios (i.e., three configurations of *TP*) strictly follow the workflow detailed in Section 3.3. *RP* and PP_X have been initialized to zero category weights. The 100 web sites in clickstream have been iteratively analyzed to decide the action to undertake (bypass, block or simulate web browsing). Whereas *RP* has been updated after each real browsing of a web sites in the clickstream, PP_{Xi} has been updated only for *non-blocked* real and simulated web browsing. Regarding the latter, we also set different values for the ratio *r* between the number of simulated and real web browsing (see Section 3.3): r=0, which implies that no simulated browsing can be used to protect the profile and, therefore, the only countermeasure is to block tracking on real browsing, and r={1, 2}.

For the three scenarios and the four values of the parameter *r*, we measured the profile protection performance of our system by measuring the divergence between *RP*, PP_X and *TP* after each iteration. To compare profiles, we used the Euclidean distance, which is a standard measure to quantify the divergence between discrete distributions [33].

$$D(P_1, P_2) = \sqrt{\sum_{i=0}^{n}((v_i;P_1)-(v_i;P_2)^2}$$,

where P_1 and P_2 are the two profiles to be compared. We preferred this distance to other measures employed in the literature on user profiling (such as the KL-divergence [14]) because: i) the Euclidian distance is symmetric; and ii) it is defined even when category weights are zero, which happens in our system during the first iterations after the profile initialization. Specifically, we report the following distances among profiles as evaluation metrics:

- $D(RP,TP)$, which represents the divergence between the target profile and the real profile built from the web sites actually browsed by the user.
- $D(PP_{Xi},RP)$, which represents the divergence between the profile learnt by the advertising platform and the real interests of the users. To protect the user profile as stated in her privacy requirements, $D(PP_{Xi},RP)$ should tend to $D(RP,TP)$.
- $D(PP_{Xi},TP)$, which represent the divergence between the profile learnt by the advertising platform and the target profile. Ideally, $D(PP_{Xi},TP)$ should approach zero.

Figures 2, 3, 4 show the evolution of these three metrics for the three scenarios we considered (*balanced*, *opposite* and *similar TPs*), respectively, and with r={0,1,2}. To make the comparison uniform

and independent of the amount of simulated web browsing that our system may introduce, distances among profiles have been updated after browsing each of the 100 web sites in the clickstream (represented in the X axis).

Figure 2. Evolution of the three metrics (profile distances) in the *balanced TP* scenario for r={0,1,2}.

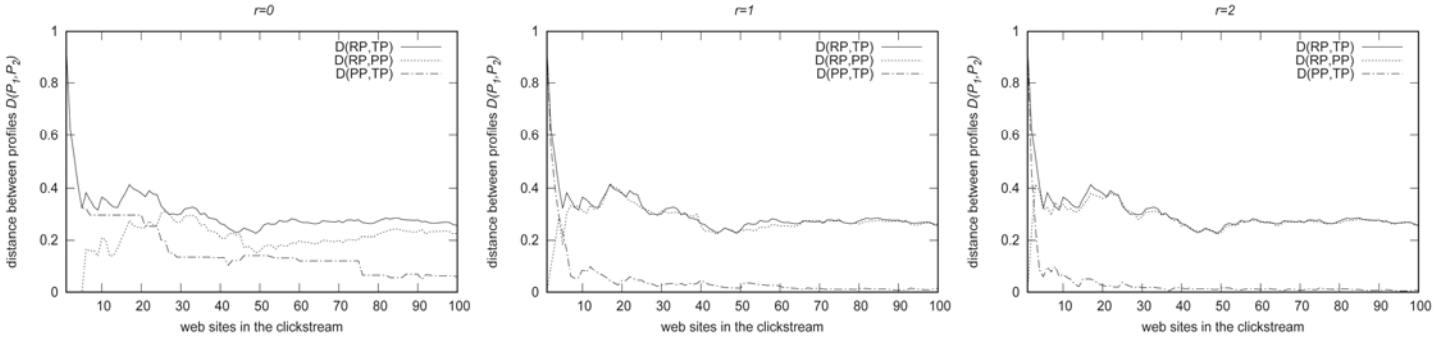


Figure 3. Evolution of the three metrics (profile distances) in the *opposite TP* scenario for r={0,1,2}.

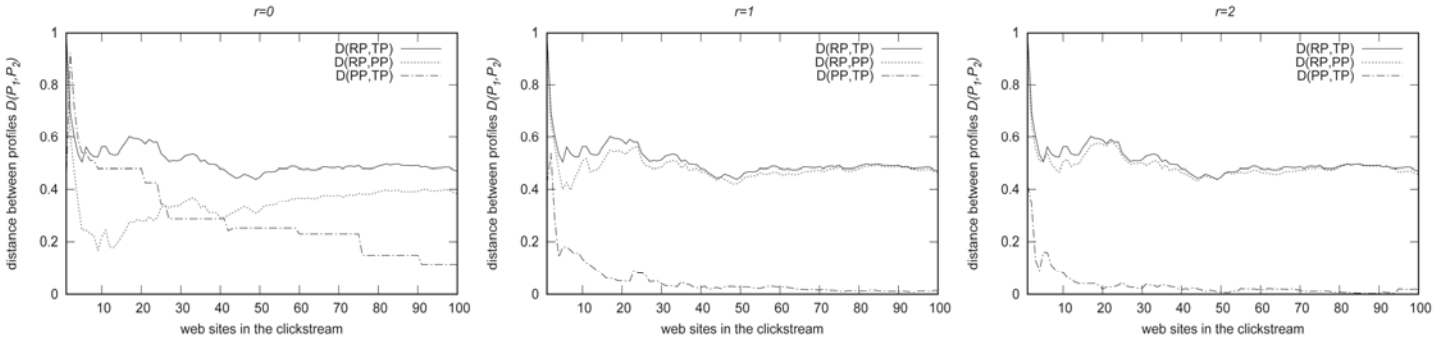
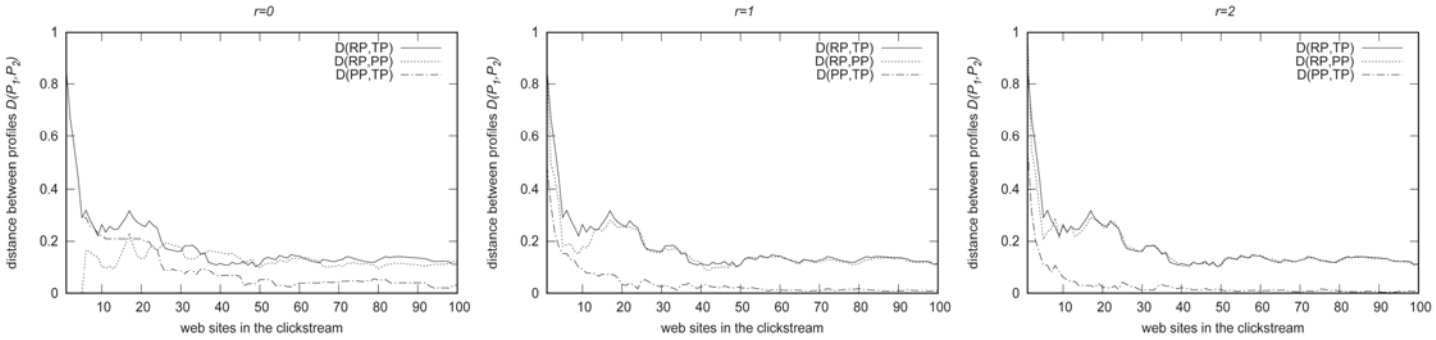


Figure 4. Evolution of the three metrics (profile distances) in the *similar TP* scenario for r={0,1,2}.



First, we can see that profiles tend to be quite stable after browsing only 20-25 web sites. Stability is reached proportionally faster to the amount of simulated web browsing introduced by our system (r) because a large r means that a larger number of web sites has been tracked by the advertising platform. The *TP* is also more accurately approximated when r is large (i.e., $D(PP_{Xi},TP)$ tends to zero, and $D(PP_{Xi},RP)$ tends to $D(RP,TP)$). Specifically, by comparing r=0 vs. r>0, we observe a noticeable improvement, both in accuracy and responsiveness when r>0; this shows the positive contribution of the simulated browsing we propose (action *A3*) in protecting the user profile. However, for r>1, the improvement is quite minor for all the scenarios, which suggests that increasing the amount of simulated browsing over the 1:1 ratio may not be worthy, e.g., because of the bandwidth overhead it causes.

The system is also able to cope with the different scenarios with similar performance. Figure 5 offers a more direct comparison of the system's performance (summarized by the $D(PP_{Xi},TP)$ metric) with respect to the value of r for the three scenarios. For the *opposite TP* scenario, we observe the largest divergence between PP_{Xi} and *TP* during the browsing of the first 15 web sites, whereas for the *similar TP* scenario the differences are the smallest. In any case, once the profiles become stable, there are not significant differences between the three scenarios for r>0: $D(PP_{Xi},TP)$ is approx. zero in all cases.

Figure 5. Top: comparison of the system's performance ($D(PP_{Xi},TP)$) with respect to r for the three scenarios. Bottom: histograms showing the distribution of action types (*bypass*, *block* and *simulate* web browsing) enforced by the system for the same parameters.

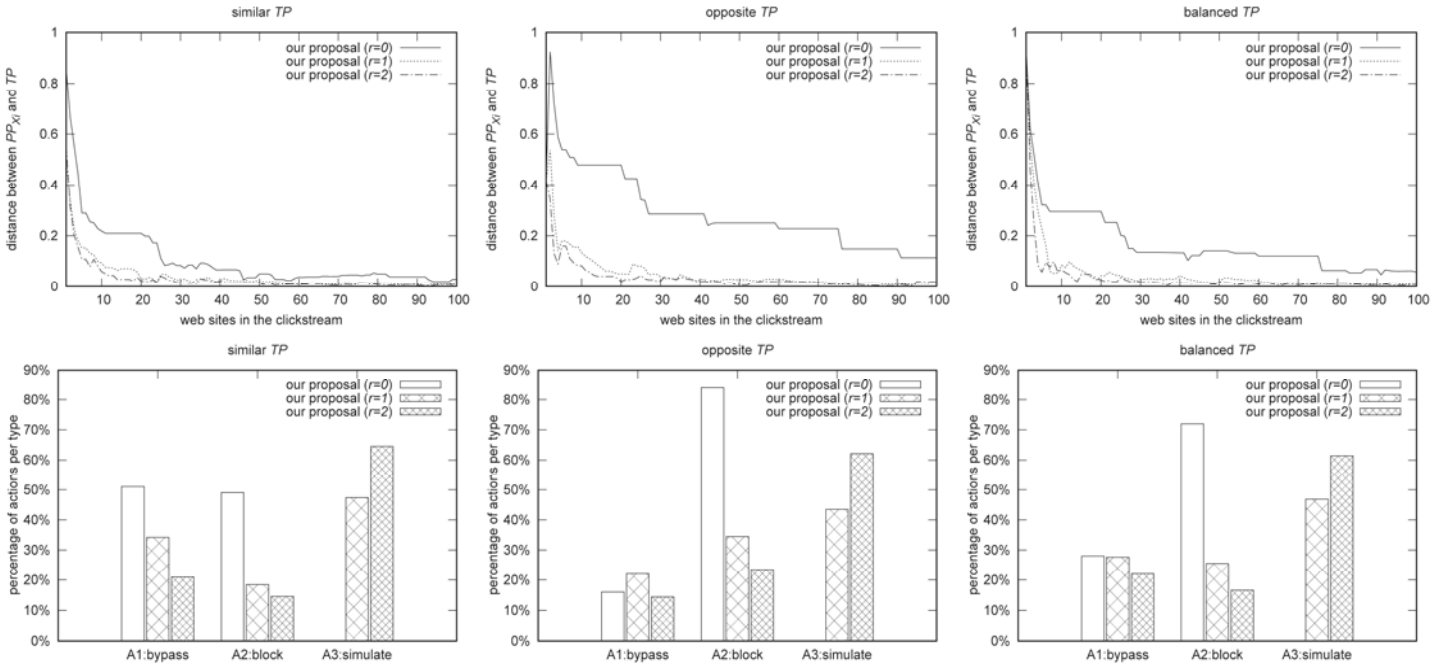


Figure 5 also reports the relative distribution of action types (as detailed in Section 3.2: *A1*, bypass a real web browsing, *A2*, selectively block tracking on a real web browsing, and *A3*, selectively simulate a web browsing) that our system has enforced for each scenario and value of r. For r=0, our system blocks the largest number of web sites. As a result, the advertising platform will gather the smallest amount of evidences on the user's interests. This amount of blocked web sites decreases proportionally to the increase of r, at the cost of increasing the amount of simulated web browsing, which may significantly distort the profile learnt by the advertising platform. The best balance is obtained with r=1: the percentage of real web accesses that are bypassed by the system is relatively large (around 22-34%), whereas the number of blocked web sites is significantly smaller than for r=0, as it is number of simulated web browsing with respect to r>1. The differences among the three scenarios are also quite minor but proportional to the differences between *TP* and *RP*: the *similar TP* scenario results in the greatest amount of bypassed web browsing (because the real interests of the user match those of the *TP*), whereas the *opposite TP* scenario results in the largest number of blocked web sites.

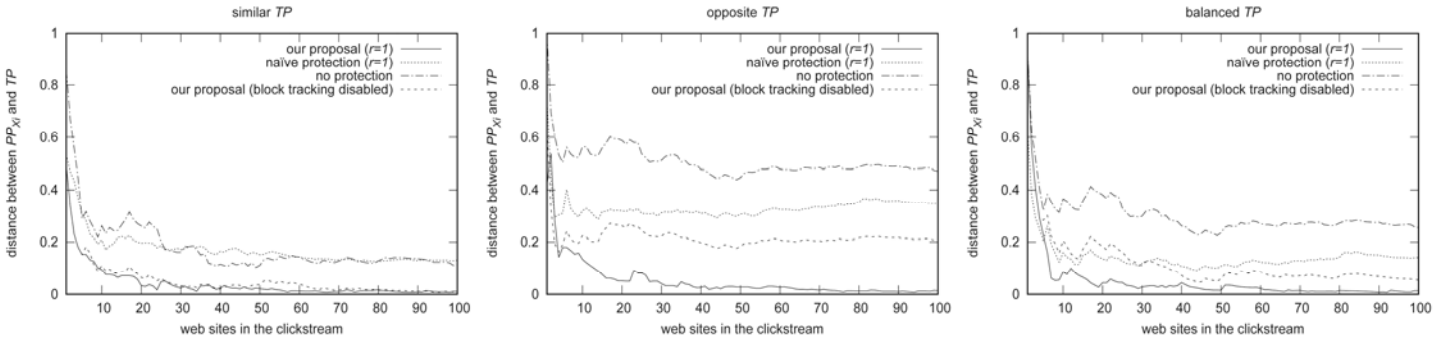
To contextualize the results obtained by our system, we compared its performance with the most balanced value of r (r=1) against the following baselines:

- *No protection* is implemented, that is, we assume that the whole clickstream has been tracked by the advertising platform.
- A *naïve protection* mechanism that tries to distort PP_{Xi} by introducing r=1 simulated browsing of random web sites uniformly chosen from the eight profile categories. Because neither profiles nor web contents are analyzed, the content of the simulated web sites will not be adapted to *TP* and simulated web sites will not be blocked. This approach emulates a naïve but widely used strategy employed to protect the users' privacy when submitting queries in a Web search engine [31].
- Our system as detailed in Section 3, but *disabling the ability to block tracking* (action *A2*). Because the only countermeasure that can be used is to execute simulating web browsing, this baseline assesses the effectiveness of action *A3*.

Figure 6 depicts the performance of our system and of the three baselines for the three scenarios. In all of them, our system is able to significantly improve the baselines by making $D(PP_{Xi},TP)$ approximate zero more accurately and faster. The *naïve protection*, which systematically executes 1 simulated browsing of a random web site is not able to adapt its behavior (i.e., the content of the random web site)

to the different scenarios. For the *similar TP* scenario, *naïve protection* is even worse than the *no protection* baseline, because the distortion it adds is uniform across the categories and, therefore, hampers the accuracy of PP_{Xi}. This uniformity makes *naïve protection* performing the best with the *balanced TP* scenario, in which the distribution of *TP* is the same as that of the random web sites (i.e., uniform). Finally, in the *opposite TP* scenario, the distortion it adds contribute to making PP_{Xi} more balanced, but not similar to *TP*. When *no blocking* of web browsing is implemented, we observe the opposite behavior: for the *opposite TP* scenario, *no blocking* performs relatively worse, because the simulated web browsing it introduces hardly compensates the real non-blocked browsing that reveals the users' interests; the best performance is obtained for the *similar TP* scenario, in which real and simulated web browsing contribute in making PP_{Xi} converge toward *TP*. In any case, we can see that the combination of the two countermeasures we proposed (i.e., selectively blocking and simulating web browsing) offers the best adaptability to the users' privacy requirements and the best profile protection performance.

Figure 6. Comparison of the performance of the proposed system ($D(PP_{Xi},TP)$) against three baselines for r=1 and the three scenarios.



5. Previous work

The work done in this field is scarce, and almost all efforts so far have focused on designing and implementing very restrictive ad blocking tools, such as *Adblock Plus* or *Ghostery*, among others; tools that mainly block all links suspicious of being related to advertising platforms. This process is done by means of pre-compiled lists of domains that are known to be linked with ad-related parties. A well-known example of those lists is EasyList[11], which is used by the Adblock tool. As it has been stressed in the introduction of this paper, while these tools effectively protect the privacy of the users, they also produce two important side-effects: i) they hamper the current Internet business model based on offering free content and services in exchange for showing advertisements and gathering personal data; and ii) they block (profile-based) advertisements that may be useful and legit sources of information about new products and trends.

Another restrictive solution that can be applied to prevent behavioral targeting is the *Do Not Track (DNT)* header [34]. This is a HTTP header field that requests that a web application to disable its cross-site user tracking. This solution does not specifically focus on blocking advertisements but on preventing user tracking instead. Nevertheless, by disabling user tracking, it also fully disrupts the behavioral targeting strategy that, in turn, results in the same issues discussed above for the ad blocking tools.

Fortunately, some researchers have gone beyond these simplistic methods and have designed systems for smart advertisement blocking. Specifically, the authors in [14] propose a technology, implemented as a web-browser extension, which enables users to configure their own access policies to enforce a smart blocking over advertising platforms. The proposed scheme is built on top of two methods that try to ascertain the extent to which user-browsing interests are exploited by advertising platforms, and that try to detect the uniqueness of the user profiles by means of a trusted central server that retrieves and compares users' profiles. The data gathered through these two methods together with the policies defined by the users allow the system to decide whether to block or not certain advertisements. Even though the authors have made an important effort to provide a smart method, which is significantly superior to the fully-restrictive ones, this proposal suffers from some issues. Specifically, the access policies defined by the users are not flexible enough to allow them to accurately define which sort of profile they want the advertising platforms to compile. The system allows the users to select the advertisements they want to

[11] EasyList. https://easylist.to/ (last accessed: 09/07/2018)

hide based on a certain set of categories and whether they are interest-based (i.e., ads are shown because they fit the user's interests), non-interest based (i.e., ads are shown because location or context fit) or retargeted (i.e., ads coming from advertisers that have been previously visited by the user); however, these categories do not put the focus on how the users desire to be profiled, which constitutes the main privacy threat. Last but not least, the system relies on a central trusted server that receives the profiles of all the users and compares them in order to compute the profile uniqueness, a measure that guides the final decision about blocking or not a certain web access. Requiring this kind of infrastructure clearly complicates the use of this solution in a real scenario, first, because of the security and privacy issues that it brings to the table, and, second, because of the lack of motivation for any organization to host and maintain a resource such as this one.

The authors in [20] propose a solution in which the user's browser is extended with a plugin to locally process the history database to determine the user's interests. In this way, the interest categorization is continually updated and refined as the user browses more web pages. Later, when the user visits a web site, the advertising platform detects whether the browser is equipped with the proposed plugin and, if this is the case, it sends to the plugin a list of possible advertisements that will be finally selected by the browser itself according to the interest categorization. Again, the first shortcoming of this proposal is that it does not take into account how the users desire to be profiled. Moreover, this solution requires certain modifications to be implemented at the advertising platform side in order to follow the proposed privacy-preserving protocol. This is a requirement that may not be accepted by the advertising platforms.

Other similar approaches in the literature that use the interests of the users to tailor the advertisements to be shown are [35-37]. These three solutions put the focus on different aspects of the same problem by means of addressing scalability issues or providing formal models and methods to help in the design of a prototype and analyzing the privacy disclosure. However, these proposals, as well as the other two solutions previously analyzed, do not put the focus on how the users desire to be profiled, which, as explained, represents a relevant privacy issue.

A different scheme that deserves attention and that is closely related to the topic being tackled in this paper is presented in [3]. This proposal evolves from [14] and presents a web-browser extension that blocks or allows advertisements depending on the economic compensation provided by the adverting platforms to the users. In other words, the authors propose a new advertising business model in which users receive economic rewards for exposing their profiles by means of the advertisements they have access to. In this scheme, the level of perturbation applied to the data and, hence, the strength of the data-protection method will depend on the economic compensation offered by each party. This solution shares with the literature, and more specifically with [14], the issue that it does not put the focus on how the users desire to be profiled; moreover, two additional difficulties appear: i) the authors assume the existence of a reliable infrastructure that allows the interaction between users and advertising platforms and permits them to negotiate the aforementioned economical compensation; and ii) the acceptance of this scheme by already stablished advertising platforms is unlikely, because it would require them to modify their business model and to adapt their infrastructure and procedures in order to work under the new rules.

Finally, it is worth mentioning that the strategies followed by the advertising industry to enforce behavioral targeting are similar to the methods implemented by web search engines (WSEs) to improve search results and to derive economic benefits from users' data. In both cases, companies gather personal data from the interaction with their users (i.e., WSEs obtain these data from the search queries that the users sent, which are assumed to reflect their interests) and, then, they compile user profiles that are used to tailor search results according to the users' interests but, also, to place targeted ads. Focusing on the protection of user profiles with respect to WSEs, because search queries cannot be blocked without impairing the search functionality, the most usual mechanism to distort user profiles consists on submitting fake queries to the WSE, an approach that bears certain resemblance with one of the counter measures we use in our proposal. More specifically, in the literature we can find simple systems such as TrackMeNot [31], which are based on periodically generating and submitting random fake queries that add noise to the profiles compiled by the WSEs. The randomness that these fake queries add to the profiles helps to protect the privacy of the users but they also disrupt their usability, which in turn hampers the accuracy of the search results. In contrast, other works have followed smarter approaches in which fake queries are tailored (by means of semantic analyses and profiling techniques) so that the profiles built by the WSEs are still useful. Specifically, the approaches in [38, 39] consider the *semantic distance* between legitimate and fake queries, so that fake queries are correlated (up to a user-defined degree) with the interests of the user. A more recent approach [32] creates fake search queries consistent with the local profiles built from the users' social network accounts (e.g., Twitter). The idea is that, while users may consider search queries private, they feel comfortable with the information they consciously

make public in their social networks. So, fake queries built from the latter bias the profiles compiled by the WSEs toward the aspects of the user's personality that are less sensitive.

6. Concluding remarks

In recent years, the advertising industry has shifted its business towards the Internet up to a degree that it currently sustains many online services open and free. However, the radical reaction of the users to protect their privacy in front of behavioral targeting -with tools that systematically block all links suspicious of being advertisements- endangers the sustainability of the current Internet model.

To tackle this issue, in this paper we have proposed a new system that conciliates users' privacy during web surfing and the advertisement-based business model in which the Internet relies. By means of semantic analyses and profiling techniques, our approach selectively blocks or bypasses tracking consistently with the user's privacy requirements and simulates browsing sessions on web sites with certain contents as a privacy-preserving measure. As shown in the empirical experiments on a variety of scenarios, the use of this simulated browsing, which is novel in the literature related to this topic, allows the system to limit the amount of blocking operations needed to preserve the user's privacy.

Our proposal also empowers the user in the protection of her privacy by allowing her to intuitively specify which of her interests can be shared with the advertising platforms. With this, we tackle several data protection goals included in the new GDPR, i.e., *data minimization* (preventing advertisers from collecting more personal data than necessary for the purpose of the service), *transparency* (users can understand which data are collected) and *intervenability* (users are given control on the protection of their data) [40]. Finally, in comparison with the naïve blocking implemented by current tools, our approach not only allows advertising platforms monetizing users' data (and, therefore, sustaining the Internet), but also retains the utility of the ads shown to the user.

As future work, we plan to evaluate our system with end users in real scenarios. With this, we will be able to gather feedback on the usability of the system and on its ability to cope with heterogenous privacy and utility needs.

Acknowledgements and disclaimer

This work was partly supported by the European Commission (project H2020-700540 CANVAS), and by the Spanish Government (projects TIN2014-57364-C2-R Smart-Glacis, TIN2016-80250-R Sec-MCloud, and TIN2015-70054-REDC Red de excelencia Consolider ARES). The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of UNESCO.

References

[1] W. Richter, Online ad revenues are surging, but 2 companies are getting most of the spoils, Business Insider, 2017.
[2] H. Beales, The Value of Behavioral Targeting, Network Advertising Initiative, 2010.
[3] J. Parra-Arnau, Pay-per-tracking: A collaborative masking model for web browsing, Information Sciences, 385–386 (2017) 96–124.
[4] D. Sánchez, A. Viejo, Personalized privacy in open data sharing scenarios, Online Information Review, 41 (2017) 298-310.
[5] T. Danova, For Mobile-Social Apps, Advertising Is Winning As The Money-Making Revenue Model, Business Insider, 2014.
[6] I. Redondo, G. Aznar, To use or not to use ad blockers? The roles of knowledge of ad blockers and attitude toward online advertising, Telematics and Informatics, In press (2018).
[7] M.R. Bauer, Are Ad-Blockers Saving Internet Users, or Ruining the Internet?, Motherboard, 2017.
[8] J.L.M.d.l. Iglesia, J.E.L. Gayo, Doing business by selling free services, Web 2.0 The Business Model, Springer, 2008.
[9] A. Griffin, Here's why you should delete AdBlock right now, Independent, 2015.

[10] S. Blanchfield, The state of the blocked web 2017 - Global Adblock Report PageFair, PageFair, 2017.
[11] R. Cookson, Google, Microsoft and Amazon pay to get around ad blocking tool, Financial Times, 2015.
[12] Wired, How WIRED Is Going to Handle Ad Blocking, https://www.wired.com/how-wired-is-going-to-handle-ad-blocking/, 2016.
[13] T.B. Lee, Google is the internet's largest ad company. So why is it building an ad blocker? , Vox, 2017.
[14] J. Parra-Arnau, J.P. Achara, C. Castelluccia, MyAdChoices: Bringing Transparency and Control to Online Advertising, ACM Transactions on the Web, 11 (2017).
[15] M. Smith, Targeted: How Technology Is Revolutionizing Advertising and the Way Companies Reach Consumers AMACOM, New York, 2014.
[16] P. Eckersley, How unique is your web browser?, 10th international conference on Privacy enhancing technologies (PETS'10), 2010, pp. 1-18
[17] M. Aly, A. Hatch, V. Josifovski, V.K. Narayanan, Web-scale user modeling for targeting, 21st International Conference on World Wide Web (WWW'12), 2012, pp. 3-12.
[18] B. Liu, A. Sheth, U.Weinsberg, J. Chandrashekar, R. Govindan, AdReveal: Improving transparency into online targeted advertising, 12th ACM Workshop on Hot Topics in Networks (HotNets'13), 2013.
[19] S. Pandey, M. Aly, A. Bagherjeiran, A. Hatch, P. Ciccolo, A. Ratnaparkhi, M. Zinkevich, Learning to target: What works for behavioral targeting, 20th ACM international conference on Information and knowledge management (CIKM'11), 2011, pp. 1805-1814
[20] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, S. Barocas, Adnostic: Privacy preserving targeted advertising, 17th Annual Network and Distributed System Security Symposium (NDSS'10), 2010, pp. 1-21.
[21] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, Z. Chen, How much can behavioral targeting help online advertising?, 18th international conference on World wide web (WWW'09), 2009, pp. 261-270.
[22] D. Godoy, A. Amandi, User profiling in personal information agents: a survey, The Knowledge Engineering Review, 20 (2006) 329-361.
[23] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28 (1972) 11-21.
[24] P. Resnik, Using information content to evalutate semantic similarity in a taxonomy, in: C.S. Mellish (Ed.) 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, 1995, pp. 448-453.
[25] S. Ross, A First Course in Probability, Macmillan1976.
[26] F. Abel, Q. Gao, G.-J.-. Houben, K. Tao, Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web, 8th Extended Semantic Web Conference, Springer, Heraklion, Crete, Greece, 2011, pp. 375-389.
[27] A. Viejo, D. Sánchez, J. Castellà-Roca, Preventing Automatic User Profiling in Web 2.0 Applications, Knowledge-Based Systems, 36 (2012) 191-205.
[28] M. Batet, S. Harispe, S. Ranwez, D. Sánchez, V. Ranwez, An information theoretic approach to improve semantic similarity assessments across multiple ontologies, Information Sciences, 283 (2014) 197-210.
[29] R.L. Cilibrasi, P.M.B. Vitányi, The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering, 19 (2006) 370-383.
[30] D. Sánchez, L. Martínez-Sanahuja, M. Batet, Survey and evaluation of Web search engine hit counts as research tools in computational linguistics, Information Systems, (in press) (2018).
[31] D.C. Howe, H. Nissenbaum, TrackMeNot: Resisting surveillance in web search, Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society, 23 (2009) 417-436.
[32] A. Viejo, D. Sánchez, Profiling Social Networks to Provide Useful and Privacy-Preserving Web Search, Journal of the Association for Information Science and Technology, 65 (2014) 2444–2458.
[33] S.-H. Cha, S.N. Srihari, On measuring the distance between histograms, Pattern Recognition, 35 (2002) 1355-1370.
[34] R.T. Fielding, D. Singer, Tracking Preference Expression (DNT), W3C Candidate Recommendation, 2017.
[35] S. Guha, B. Cheng, P. Francis, Privad: Practical privacy in online advertising, Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation (NSDI 2011)Boston, MA, USA, 2011, pp. 169-182.
[36] H. Haddadi, P. Hui, I. Brown, MobiAd: Private and scalable mobile advertising, Proceedings of the 5th ACM International Workshop on Mobility in the Evolving Internet Architecture (MobiArch 2010), ACM, Chicago, IL, USA, 2010, pp. 33-38.

[37] Y. Liu, A. Simpson, Privacy-Preserving Targeted Mobile Advertising: Formal Models and Analysis, International Workshop on Quantitative Aspects in Security Assurance, 2016, pp. 94-110.
[38] D. Sánchez, J. Castellà-Roca, A. Viejo, Knowledge-based scheme to create privacy-preserving but semantically related queries for web search engines, Information Sciences, 218 (2013) 17-30.
[39] B. Shapira, Y. Elovici, A. Meshiach, T. Kuflik, PRAW—A PRivAcy model for the Web, Journal of the American Society for Information Science and Technology, 56 (2005) 159-172.
[40] North Rhine-Westphalia State Commissioner for Data Protection and Freedom of Information, Standard Data Protection Model, 2016.