Computer Networks 57 (2013) 2833–2855

Contents lists available at SciVerse ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Privacy leakage on the Web: Diffusion and countermeasures ☆

Delfina Malandrino *, Vittorio Scarano

ISISLab, Dipartimento di Informatica, Università degli Studi di Salerno, I-84084 Fisciano (SA), Italy

ARTICLE INFO

Article history:
Received 9 November 2011
Received in revised form 8 April 2013
Accepted 13 June 2013
Available online 6 July 2013

Keywords:
Online privacy leakage and threats
Privacy enhancing technologies
Web navigation

ABSTRACT

Protecting privacy on the Web is becoming increasingly complicated because of the considerable amount of personal and sensitive information left by users in many locations during their Web browsing and the silent actions of third party sites that collect data, aggregate information and build personal profiles of Internet users in order to provide free and personalized services. On the other hand, most of people are unaware that their information may be collected online, and that, after their aggregation from multiple sources, could be used for secondary purposes, such as linked to allow identification, without user's notice.

We present, in this paper, an empirical data study in order to describe how users' privacy can be undermined because of a variety of potential privacy threats on the Web, mainly perpetrated by third party entities against unaware users, and to quantify the penetration of these third party domain servers into their online activities. Moreover, we discuss our methods and findings to protect the individuals against invasions of their privacy and to limit the diffusion of personal and sensitive information during Web browsing. Specifically, we present a supportive, comprehensive and improved approach for privacy protection to allow users to be *aware* of the risks of their navigation and to give them *full control* on feasible actions to address the risk of several privacy threats. We envisioned a comprehensive approach to face privacy leakage by adding to the traditional URL-based filtering mechanism a new filtering method which allows to address privacy threats unprecedentedly not dealt with. Our approach is validated by a Firefox extension, named NoTrace, that brings together several existing techniques in this field but also implements new improved techniques that ensure better privacy protection. We used NoTrace to broadly analyze the Web in order to inspect the potential threats contained in the most popular Web sites and inform online users about both their risk and extent. This data set was also used to test the efficiency of NoTrace for effectiveness and performances which allows us to mark a definite improvement on privacy protection for users while navigating the Web.

1. Introduction

The increasing number of daily activities that can be performed on the Web causes an increasing risk for Internet users about the privacy of their personal information. Online browsing, online banking, online transactions, online shopping, social network interactions and any type of online collaboration and communication could undermine the privacy of the individuals because of the frightening increase of the Web information leakage, especially when they visit popular Web sites. Specifically, the users' information can be accessed, gathered, stored, data mined, linked, shared, contracted and potentially sold, for profit-making purposes, and mainly, *without permission or consent*. The data trails and the footprints left in many locations during Web browsing allow third party

☆ A preliminary version of this work appeared in Poster Proceedings of the Third IEEE International Conference on Privacy, Security, Risk and Trust [1].

* Corresponding author. Address: Dipartimento di Informatica, Università di Salerno, Via Ponte don Melillo, I-84084 Fisciano (SA), Italy. Tel.: +39 089 969714; fax: +39 089 969600.

E-mail addresses: delmal@dia.unisa.it (D. Malandrino), vitsca@dia.unisa.it (V. Scarano).

1389-1286/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comnet.2013.06.013

companies and aggregators to build digital dossiers of the Internet users, by threatening the privacy of both their online and private life, and raising a number of interesting privacy issues.

There have been many attempts to define privacy, from synonymous with the "right to be let alone" [2] to the right to prevent the disclosure of personal information [3]. Moreover, tied to the definition of privacy, there is also the definition of privacy concern. Several studies have consistently proved that the majority of people are concerned about threats to their privacy when they are online [4–9] and that privacy concerns influence people's willingness to disclose personal information to a Web site [10]. However, studies focussing on the relationships between Internet users' concerns and the corresponding actions to take, discovered an apparent dichotomy between privacy attitudes and resultant behaviors [11,12], highlighting situations in which users are less willing to take actions to protect privacy (and, therefore, more likely to share their personal information and preferences) when some benefits can be obtained in return [13,14]. In fact, studies in this field show that only a small percentage of users read privacy policies [15] and, in general, most of the users is not able to reliable understand their content [16,17]. Additionally, users show little willingness to adopt privacy protective technologies [11].

In recent years, it is widely accepted that the online marketing methods of network advertisers have given rise to concerns about users' privacy [18,19]. Large business companies, as anticipated before, often intervene as uninvited guests during the Internet experience, by monitoring users' behaviors, actions and habits to provide them targeted advertising [20–22]. The practice of tracking individuals' online activities is not a dangerous activity *per se*, and although it increases the effectiveness of the marketers' campaigns and also their revenues,[1] it also undermines users' privacy since their data, collected and aggregated, could be used by third party entities for "*potential*" malicious activities, instead of what the behavioral advertising has been envisioned for. It is worth to note that behavioral advertising is expected to continue its double-digit yearly growth through 2014 as Americans increasingly go online [18].

The main problem is not strictly related with the information collection, but about the final use made by third party companies of these data. Therefore, the concern is about the possibility that these pseudo-anonymous data, aggregated and linked with personally identifiable information (i.e., email addresses, full name, address, phone number, fax number, credit card number, social security number, etc.), may be disclosed, or explicitly sold by first sites, to third party entities and potentially used for secondary activities, such as identity theft (the fastest growing crime in US), social engineering attacks, online and physical stalking and so on[2,3] [24]. Personally identifiable information can be also used to identify individuals not only in future tracking but also retroactively by using data that's already been collected [25].

It is important to emphasize that privacy advocates criticize behavioral advertising because it results in the compilation of a sizable array of potentially sensitive data about an individual that often is unaware of his/her leakage and is not able to monitor, protect and control it [24,19,26–28]. As stated in [29] "*Indeed, profiling arguably harms individuals regardless of how it is used because it results in an unprecedented loss of privacy*". And, moreover, "*By merely participating in the Internet economy, consumers lose control over which details about their private lives are known, and they have little control over who gets to learn of these details after the data passes into a profilers hands*" [29].

Among the existing regulations, the Canadian legislation (4.9 Principle 9) states that *"Upon request, an individual shall be informed of the existence, use, and disclosure of his or her personal information and shall be given access to that information"*.[4] The Directive 95/46/EC of the European Parliament and of the Council (Article 7) states that collect and process personal data of an individual is not legitimate if those activities happen without his or her consent.[5] Ultimately, users cannot assess the potential magnitude of harm because they do not know when and in which extent profilers are collecting or guessing data about them. Opportune countermeasures have to envision how to make users aware of their information leakage during online activities, and give them full control over the potential improper use of their personal and sensitive information.

In a series of significant studies, information leakage has been recognized as a pervasive problem [30–33]. More recent statistics show that 63% of users agreed with a statement of concern for third parties monitoring activities [34] while another study [33] has highlighted the criticality of the problem showing that 56% of sites analyzed (75% when considering userids) directly leak sensitive and identifiable information to third party aggregators. The widespread of availability of extensive records of individual behaviors, the merging of the anonymous clickstream of data with personally identifiable information and their disclosure to third party sites, without permissions or consents, present several challenges related to privacy raising concerns among Internet users that call for more effective solutions to get protection against invasions to their privacy.

While the ability of network advertisers and third party companies/aggregators to collect an increasingly amount of personal information about users when they are online has been steadily growing, the awareness about the persistent consumption of the users' privacy is growing slowly [33]. Users need to be firstly informed about the risk of some activities on the Web, and secondly, educated about their privacy practices. To achieve this goal specific countermeasures to address privacy leakage, and tools able to guarantee awareness and user control, friendliness, comprehensiveness, effectiveness and performance, have to be envisioned and developed.

[1] In the first half of 2012, internet advertising revenues climbed to an all-time high of $17 billion according to IAB Internet Advertising Revenue Report [23].

[2] http://www.priv.gc.ca/media/nr-c/2012/nr-c_120925_e.asp.

[3] http://www.idtheftcenter.org/ITRCBreachReport2013.pdf.

[4] http://laws-lois.justice.gc.ca/PDF/P-8.6.pdf.

[5] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:NOT.

It is worth saying that several work exist in this field [31,32,35,36,7,38–41] and several techniques [42] and tools, both intermediary-based [43,44] and client-based [45–48], have been designed and implemented. In addition, practitioners, researchers, and advocates have begun to more formally study how Web sites collect, use, and share information about individuals, since as discussed before, data gathering, and mainly their misuse, pose significant privacy problems, causing increasing concerns among Internet users[6] [29].

In summary, any approach that aims to achieve effective and meaningful results in the battle against the loss of privacy has to consider both the criticality of the personal information disclosure (when disclosure is without consent) and its corresponding proposed or advertised use (often not clear, and misleading in most cases), and, at the same time, it has to enhance both awareness about privacy risks and full control on the feasible actions to take, without interfering with the use of the Web.

1.1. Our contributions

We present in this paper an empirical study aiming at analyze the diffusion of some potential threats that could affect the privacy of users during their Web navigation, and the role played by third party entities when we analyze the extent of the corresponding information leakage.

A broad study of the Web, through the analysis of most popular Web sites, can help users to increase their awareness about information leakage, researchers to study new measures to help individuals to guard against harmful activities, and finally, developers to provide new and more effective instruments to fight the battle against the deceitful activities of third party aggregators and, more recently, of first party sites that should (but do not) safeguard individual privacy [33].

To accomplish this study we have used NoTrace[7] [1], a Mozilla Firefox extension, whose main goal is to provide several instruments to protect privacy of users during Web navigation and to limit the diffusion of sensitive and personally identifiable information. We envisioned, through the NoTrace implementation, a new supportive, comprehensive and improved approach for privacy protection, in order to make users aware of the risks of their navigation and to give them full control on feasible actions to address them.

Our approach takes into account the traditional "*URL*"-based blocking technique (used by the popular Adblock Plus [45] and NoScript [46] Firefox extensions) and also our new "*Content*"-based filtering technique, that is accurately placed (in the processing pipeline of Mozilla Firefox browser) even before the browser rendering begins, therefore avoiding a large variety of loopholes and tricks that could be used to overcome the filtering.

NoTrace allowed us to witnesses the feasibility and to measure the *improvements* of our approach for privacy protection over a wide data set, with respect to other tools, as well as proving that effectiveness of the techniques does not come at the expenses of unacceptable loss of performances. The comprehensive performance study that we have realized and documented in the second part of the paper shows, indeed, that the high quality of user's experience, the efficiency of his/her navigation and the minimization and control over privacy loss are a reasonable and obtainable joint objective.

The organization of the paper is as follows. Section 2 introduces the main motivations of our work. Section 3 discusses a classification of potential threats that could affect the privacy of the users during Web browsing. Section 4 explores the weakness of existing systems for privacy protection on the Web and how our approach can overcome them. In Sections 5 and 6, that present the methodology and the study results, respectively, we show results about the extent of the privacy leakage, as well as the NoTrace improvement on privacy, the effectiveness of the implemented techniques, and finally, their minimum impact on user performances and experiences. In addition, we compare NoTrace with other popular Firefox extensions in this field. Next, in Section 7 we present related works while in Section 8 we conclude with some final remarks.

2. Motivation

In this section we introduce the main motivations of our work. First and foremost, we emphasize the need of an holistic view to study the potential risks to which users may be exposed during their online activities and to study the corresponding countermeasures.

Different empirical studies, available in literature, have analyzed the risk of many privacy threats [8,38,39,49–51]. Specifically, in each study authors focussed on a specific threat, performing, therefore, a targeted analysis and conveying results that were strongly related to environment settings and to the data sets specifically tested at the moment of their experimental studies. It is worth to note that, if a general view is required, a comparison among these separate studies is not informative, even if they showed a common objective. Our goal is to perform a comprehensive experiment, that involves the simultaneous analysis of many potential privacy threats, tested in the same environment (same hardware and software), at the same time and on the same data set.

The second motivation that prompted us to undertake this work was the weaknesses of the existing systems and approaches to protect privacy on the Web (whose detailed discussion is presented in Section 4.1). First of all, the missing *support* for users when informed decisions are needed to protect the own privacy while navigating the Web. Second, the need to address in a comprehensive way many privacy threats, to avoid for example, to install and configure several, sometimes conflicting extensions into the browser. Awareness and full control, considered very important requirements by the Federal Trade Commission (FTC) [52], the Privacy Right Clearinghouse (PRC) [20] and other important privacy watchdog organizations, have not yet been envisioned by existing privacy tools. Moreover, several tools in this field provide standard

[6] May 2012 Web Privacy Measurement, http://www.law.berkeley.edu/12633.htm.

[7] http://www.isislab.it/projects/NoTrace/Download/Beta/notrace@unisa.it.xpi.

measures for privacy protection without distinguish between different attitudes, beliefs, and technical experience of the users, and therefore, without envisioning the possibility to convey information in a way that is understandable for all target users.

To adhere to the recommendations discussed earlier, and overcome the limitations of the state-of-the-art existing systems, we have designed a new approach (that we will discuss in Section 4) for privacy protection that, while addressing efficiently a large number of privacy threats (as the next two sections will substantiate), does make users aware of their privacy leakage, and inform them when it happens and with which extent, by giving them full control on the possible actions to take.

3. Online privacy threats

Many systems exist in literature to preserve online privacy or to disable Web tracking [53,45–47,54,55,7,56–58]. Unfortunately, these systems are capable of protecting users by addressing only some specific issues. For example, NoScript disallows JavaScript, Java and other executable content and Web tracking services that rely on active client-side content. Adblock Plus mainly filters out annoying advertisements. It supports filters that allow users to set rules manually about objects that have to be blocked or allowed (with objects that must be expressible in Adblock filter language). Its EasyPrivacy filter subscription,[8] specifically designed to protect the privacy, stops tracking content, such as Web bugs and unwanted JavaScript files. These extensions are very popular[9] but they are not capable to ensure, alone, full privacy protection.

Moreover, since they address different and separate issues, a user should install several extensions at once, with some conflicts easily arising among policies and actions, with the overall consequence to have a general slowdown in performances, as we shown in our performance study described in Section 6. Firefox Mozilla is known to reduce performances where several extensions are active at the same time.[10]

3.1. Classification

In this section we introduce a classification of some online privacy threats, discussing in detail the troubles that they can involve during users' online activities.

3.1.1. Personal information

3.1.1.1. Leakage of browser environment information. As part of the HTTP exchange browsers pass environment variable data that include, for example, information about host, Operating System, browser type, user's preferred language, details about installed Plugins, system fonts and so on. The risk here is about the client identification through browsers characteristics, also known as "*browsers fingerprinting*" [59].

Referer: Piece of information extracted by the HTTP request that could be used to infer users' movements and habits [33]. They can also leak sensitive information (e.g., search terms or user IDs) [60].
User Agent: Piece of information extracted by the HTTP request to infer user browser type and Operating System information. This information can be used for different privacy attacks [61].
From: Piece of information extracted by the HTTP request to infer the Internet e-mail address of the user issuing the request. Leakage of this private information occurs, for example, after an account confirmation [33].
IP Address: Piece of information used to perform Geo-Targeting and serve advertisements to users according to their geographic location. Users may not be aware of what else their location is used for (e.g., Behavioral advertising). Although location-based services involve potential advantages, the proliferation of devices that geolocate themselves raises dramatic concerns about users privacy [62].

3.1.1.2. Leakage of personally identifiable information. It includes personally identifiable information about users inspected through the analysis of histories or through navigated URLs [33]. It is worth to note that a well-known result in linking pieces of personal identifiable information is that most Americans (87%) can be uniquely identified from a birth date, zip code, and gender information [63].

HTTP Cookies: Tiny files placed on users computers to help Web sites to recognize users. Cookies are also used to track users through history data saved inside them [9].
Browser History and Auto-Complete: The browser history allows Web sites to inspect users personal information from their actions and behaviors, and can be used to de-anonymize social network users [64]. Auto-Complete is a feature provided by browsers to provide, when typing a new request, a list of previously visited Web sites.
Identifying URLs: URLs that pass user personal information (i.e., email addresses, full name, address, phone number, fax number, credit card number, social security number, etc.), through query strings, to Web sites. These URLs contain character code such as '?', '=', or '&'. Elimination of these URLs reduces the capability of information to be passed as part of the URL, although it could remove needed content for the page [42].
Script executions: JavaScript and general script executions are potentially a privacy concern as the code can gain access to browser information, such as cached objects and the history of visited links [37,38,61].

3.1.2. Web tracking

It concerns the, typically invisible, practice of tracking individuals to analyze behaviors and habits and suggest them tailored advertising.

Web bugs: Small GIF images (1×1 pixel size) inserted in the HTML code of the Web page to track users'

[8] https://easylist-downloads.adblockplus.org/easyprivacy.txt.
[9] Adblock Plus is the most popular Firefox add-on with 15,297,038 users and NoScript is the second most famous with 2,061,712 users. Data retrieved on 5th April, 2013.
[10] http://blog.mozilla.com/addons/2010/06/14/improve-extension-startup-performance/.

movements. Through Web bugs it is possible to inspect the following information: IP address, URL of the Web page hosting the image, time when the image has been viewed, the browser type and `Set-Cookie` values. Web bugs can transfer previously inserted personal information (i.e., username, password, email address) [65]. Web images present also privacy concerns since they leak information via Identifying URLs [42].

Third party cookies: Cookies extensively used by advertising companies when the user is visiting a Web page that contains third party content from third party servers [33,66].

Flash Cookies: Flash-cookies [49], also known as Local Shared Objects (LSOs), are pieces of information placed on computers with the Adobe Flash browser plug-in and allow to customize settings for Flash-based applications (i.e., Youtube video player). Those cookies are placed in central Adobe's system folders and so protected from deletion (they cannot be erased through the standard cookie privacy controls in a browser). They are capable to store 100 KB's of information for an indefinite amount of time. To delete Flash cookies a user has to visit the Adobe's Web site and use the provided Global Storage Settings panel. They are also able to "respawn" or re-instantiate HTTP cookies deleted by the user.

Hidden third party objects and objects from top aggregation servers: Object retrievals that could be used by hidden (a given server looks like it belongs to a first-party domain, but actually belongs to a third party [38,39]) third party servers to aggregate information about a user's page retrieval.

Tracking from *ad*-networks: Execution of JavaScript code by *ad*-networks and the use of third party cookies to track users across many Web sites. To avoid Web sites to install third party cookies by *ad*-networks, users can accept opt-out cookies, informing that they do not want to be tracked anymore [19].

HTML5 and ETag Respawning: The HTML5 APIs Local Storage provides a way for Web sites to store and retrieve large data locally, within the client browser, through JavaScript code [67]. Because designed for larger quantity of information (the W3C recommends an initial limit of 5 MB per origin), respect to the limited size (i.e., 4 KB) used by cookies, advertisers and other third-party sites could potentially see weeks or even months of personal data, which could include a user's geographical location, time zone, photos, shopping cart contents, e-mails and histories of Web pages visited.

Since the recent nature of these APIs (published as Candidate Recommendation[11]) and their not yet widespread usage [68], techniques and/or tools to control and limit access to local storages only by authorized parties have not yet been envisioned. The existing techniques only allow users to view and delete data items stored in the local storages, even most of these entries are frequently obscure and only in specific instances will the originating Web sites and other information be understandable.

The ETag tracking [51] is particularly problematic because the technique generates unique tracking values even where the user blocks HTTP cookies, Flash and HTML5 cookies. Specifically, the information the ETags contain allow for deleted cookies to be recreated so Web tracking can continue, with no awareness for Internet users. The most threatening thing is that ETags can track users even if they have activated the private browsing mode for their browser sessions.

3.1.3. Third party activities and unwanted advertisements

It includes object retrieval from third party servers and threats posed by malicious interfaces on the Web.

Third party objects and objects from top aggregation servers: Object retrievals that could be used by third party servers to aggregate information about a user's page retrieval. All content from the domain (2nd-level DNS name) of the user is allowed while requests from other domains are filtered [38,39].

Advertisements and Malicious interfaces: They are able to trick, coerce, or manipulate users into taking undesired actions, such as, for example, disclosing personal information, making a purchase, clicking on an advertising link, signing for undesired mailing lists, forcing the user to wait and view undesired content, and so on. With these activities, designers sacrifice users time, attention and personal information, by performing (frustrating) cognitive attacks [69]. Techniques that block advertisements and unwanted content could be used to address this type of attacks.

4. Our approach

In this Section we describe the lacks of the existing tools for privacy protection and the approach we have designed and implemented to overcome these limitations.

4.1. Weakness of existing privacy tools

A lot of functionalities to protect privacy have been provided by client-side technologies such as Firefox extensions or add-ons. Specifically, we thoroughly studied Ghostery [48], Adblock Plus, RequestPolicy [47] and NoScript, the most popular add-ons listed in the "Security and Privacy" category of the Mozilla Community.[12]

We want here to describe what we believe are the main weakness of these systems when used to protect privacy on the Web.

The first weakness is the missing support for the user during his/her navigation. As a matter of fact, little or no support is given by the existing systems to the users to understand the nature of the risks to which they are continuously exposed while navigating the Web and the extent of their information leakage. They also do not provide adequate suggestions to unexperienced users to guide them into the appropriate countermeasures.

[11] Web Storage specification, December, 8, 2011, http://www.w3.org/TR/webstorage/

[12] https://addons.mozilla.org/it/firefox/extensions/privacy-security/.

All the analyzed tools rely on standard blacklist/whitelist mechanisms of privacy protection. Specifically, AdBlock Plus is structured as a blacklist, this means that URLs to block to remove online advertisements and know malware domains have to be specified. Ghostery is another example of blacklist-based system to detect trackers, Web bugs, pixels, and beacons placed on Web pages by *ad*-networks. It provides feedback about the companies and aggregators it blocks but any type of configuration or personalization has been envisioned. NoScript is structured as a whitelist in order to allow JavaScript, Java and other executable content to run only from trusted domains. It is an URL-blocking tool, but it does not provide feedback to users about the blocked scripts (it only shows the number of objects blocked). Similarly to NoScript, RequestPolicy is a Firefox extension that allows users to block cross-site requests unless they specifically choose to allow them through their specification in a whitelist. Moreover, RequestPolicy starts with a default whitelist that contains the most popular Web sites, in order to limit the number of page breaks.

Overall, these tools are ineffective at communicating their purpose and guiding users to properly configure them [70]. They do not show the real risk to which users could be exposed, and in addition, very often, default settings are passively accepted, since average users are not able to understand, without any help, how to change them, to increase the level of protection or to fix broken elements.

To tackle the largest number of privacy threats, users have to install and configure several, sometimes conflicting[13] extensions into the browser, since the most solutions in this field only address (by design) specific privacy threats. Specifically, some of them only provide information about third party objects counts (i.e., NoScript), other provides information about which companies track users movements (i.e., Ghostery), while others simply show the URLs that they block (i.e., AdBlock Plus and RequestPolicy).

To ensure privacy on the Web, we believe that an effective approach should address, the most critical privacy issues, but involve the user into "the loop" and allow him/her the comprehension of the what is going on, by ensuring two important requirements: awareness and full control. First of all, users should be able to know who is tracking their movements and gathering their information, when it happens, by which means, and with whom this information will be shared. Users should be able to decide whether and what actions to take, whether or not Web sites can share information about them with corporate affiliates or third party servers. It is important to highlight that, from the awareness point of view, any tool has yet been provided to allow users to be aware of all the information they disclose to third party servers.

Several tools in this field provide standard measures for privacy protection without distinguish between different users' attitudes, beliefs, and technical experience. For example, some users do like targeted or personalized advertisements [71], so removing them unconditionally, through specific countermeasures, may not reflect their real preferences and/or needs. Users navigation may also widely change over time, thus involving corresponding changes in their privacy settings. Moreover, users differ widely in their privacy concerns, with people concerned over unauthorized others accessing their personal data, and others concerned about the risk of secondary use, that is the reuse of their personal data for unrelated purposes without their consent (personal data sharing and aggregation activities). People also differ in their level of concern as firstly stated in [72], and they can be classified in three different clusters of users, that is, marginally concerned, privacy fundamentalists, and the pragmatic majority. This means that to meet the needs and the expectations of the broader range of users, tools and approaches to protect privacy have to address both the wide range of existing privacy threats as well as their different corresponding levels of concern.

[13] Ghostery message: *"Warning! When combined with other cookie monitoring addons such as Beef Taco, Cookie Monster, and Google Opt-Out, this feature can cause un-responsive script errors. If you experience this error, please try disabling this feature or conflicting addons"*.

4.2. Strenghts of our approach

Our approach for privacy protection, as introduced in Section 1, exhibits the following features: support for users and comprehensiveness, awareness and full control, personalization. First of all, we believe that it is crucial for not expert or novice users to know that they can leverage useful suggestions when important actions must be taken about privacy preservation, and that, at the same time, such actions do not will undermine both the functionality and the quality of the accessed Web pages.

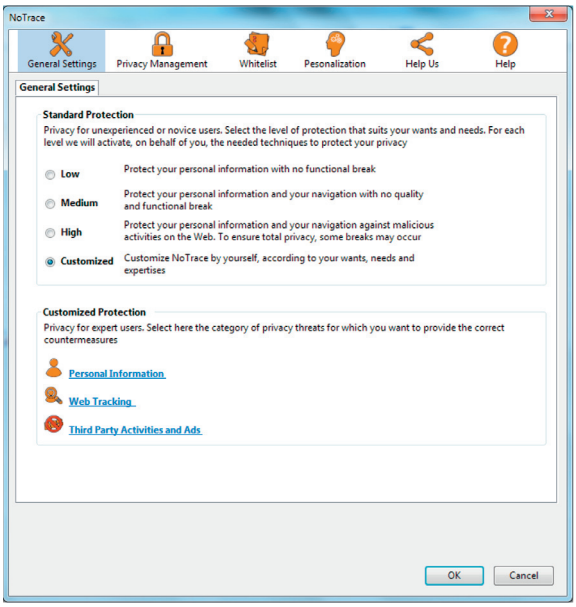
In addition to assisting users, our approach aims to make them aware of what silently happens during their navigation, by informing them about who have access to their information, and helping them to avoid their exposure to various physical and online risks. It is crucial to make extremely difficult for third party sites to continue with the inspection of users' activities whose goal is to build digital dossiers of their behaviors.

Moreover, offering personalized protection by taking into account preferences, needs, users expertise, beliefs and general attitude toward privacy, represents another requirement of our approach, while offering a single tool (that encompasses several functionalities) to control all aspects of privacy on the Web can avoid to install several systems and tools and ensure better performances.

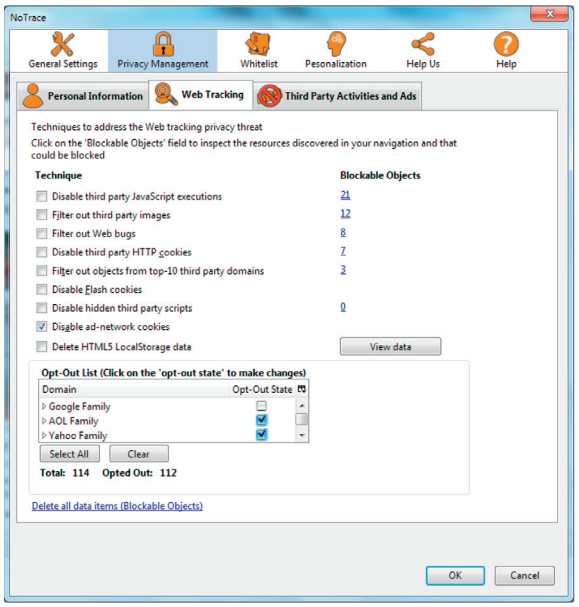
These features, introduced by our approach and embodied into NoTrace, are described in detail in the following.

4.3. Supportive privacy protection

To allow NoTrace to be supportive toward all users, who can exhibit different levels of concern and knowledge of privacy violations, we have envisioned two different modalities of privacy protection, that is, a "Standard Protection" for novice users and a "Customized Protection" for experienced users (see Fig. 1(a)). With the former modality, users can choose among three different pre-defined levels of privacy protection, "Low", "Medium" and "High", to enable techniques that are able to ensure the corresponding level of protection. Each level, therefore,



(a) Privacy for novice ("Standard Protection") and experienced users ("Customized Protection")



(b) Fine-tuning setting for the "Customized Protection" modality

Fig. 1. NoTrace configuration: general settings and privacy management panels.

provides various privacy settings, organized taking into account the different users attitudes toward privacy: privacy over quality and functionality (i.e., High level), quality and functionality over privacy (i.e., Low level), or a trade-off between them (i.e., Medium level). Our aim, here, is to provide support for unexperienced users who do not want (because not capable) to decide which actions to take but trust NoTrace and fully rely on its functionalities.

The "Customized Protection" allows advanced configurations for expert users. Users are free to choose any of the provided countermeasures for the selected category of privacy threats they want address (i.e., "Personal Information", "User Tracking" or "Third Party Activities and Ads"). Conversely to the previous modality, users can further fine-tune the privacy settings, enabling and disabling specific functionalities whenever they want. In Fig. 1, as an example, we can see how the user has selected the "User Tracking" Panel, and specifically, the technique that allow him/her to opt-out from the tracking performed by a large number of advertising companies (i.e., more than 40 companies that have joined the Network Advertising Initiative[14] initiative), excluding the Google Family from the *opt-out decision*, to access, for example, the Gmail account without errors.

4.4. Comprehensiveness

NoTrace is able to address several online privacy threats (discussed in the previous Section) by providing, all in one, opportune countermeasures, whereas each of them is singularly provided by other popular extensions in this field (that sometimes even fight each other[15]). Specifically, we have implemented all techniques shown in Table 1, whereas some of them have been discussed in our earlier work [1]. Some of these techniques leverage the new "Content-based" filtering mechanism. It acts on the content of Web pages, by parsing the corresponding HTML code to discover inline third party objects. An example is the *nojscookie* technique that looks at the very popular document.cookie DOM field [68] and filters out it.

The HTML inspection via the publish/subscribe design pattern, available via the Mozilla API, allows NoTrace to access the HTTP stream before the browser. No other tool has harnessed this type of filtering before, leveraging the URL-based filtering mechanism only.

In Table 1 we show which techniques leverage each of the envisioned filtering mechanisms (i.e., the *U* value refers to the standard URL-based blocking, the *C* value refers to the Content-based mechanism, while the *HM* (Header Management) value refers to techniques that deal with information extracted by HTTP requests/responses and the *S* (Storage) value to techniques that have access to the file system and local storages).

4.5. Awareness and full control

Users should be able to know who is tracking their movements and gathering their information, when it happens and with whom this information will be shared. To educate users on what information about their browsing behavior is sent to third-party aggregators and the information that is inferred based upon these behaviors, we have envisioned in NoTrace, the possibility to steadily monitor the Web nav-

[14] http://www.networkadvertising.org/.

[15] http://adblockplus.org/blog/attention-noscript-users.

Table 1
Description of the privacy protection techniques implemented in NoTrace.

Filtering mechanism	Technique	Description
HM	*nocookie*	Disable all cookies
HM	*no3cookie*	Disable third party cookies
S	*noflashcookie*	Disable Adobe Flash cookies
C + U	*nojs*	Disable JavaScript executions and filter out ⟨script⟩ ⟨/script⟩ HTML tags
C	*nonoscript*	Filter out ⟨noscript⟩ ⟨/noscript⟩ HTML tags
U	*no3js*	Disable third party JavaScript executions
C + U	*noimg*	Filter out Web images
U	*no3img*	Filter out third party images
C + U	*notop*	Filter out objects from Top-10 third party domains
C + U	*noad*	Filter out *ad* objects
U	*no3obj*	Filter out all third party objects
C	*nowebbug*	Filter out Web bugs
HM	*noidheader*	Filter out potentially identifying headers in the HTTP request
U	*noIdentURLs*	Block URLs that contain character code such as '?', '=', or '&' to pass user personal information to Web sites
C + U	*no3rdhiddenobjs*	Disable *ad*-networks JavaScript executions
S	*noadnetwcookie*	Accept opt-out cookies from ad networks
C	*nometacookie*	Filter out ⟨META HTTP-EQUIV="Set-Cookie"⟩ tags
C	*nojscookie* & *norefererrcookie*	Filter out JS document.cookie and document.referer function calls
C	*nometaredirect*	Filter out ⟨META HTTP-EQUIV="refresh"⟩ tags with content value set to URLs for third party resources
S	*nolocalstorage*	Show and delete (on per-user based preference) data items stored in the browser LocalStorage

igation, allowing users to be constantly informed about the third party objects discovered during Web browsing and the extent of their information leakage. Specifically, similarly to other tools in the same field, NoTrace provides a "Blocked Object Panel", shown in Fig. 2, to inspect which URLs have been blocked since requesting objects from third party domains.

NoTrace allows users to examine portions of their visited Web sites, to establish which third party objects were requested. This option is available through the "Blockable Objects" link information, as shown in Fig. 1. In that Figure we can also see that through the "Advanced" and "View Data" buttons it is possible to inspect which scripts could be used to track users movements by hidden third party domains and which information third parties can store in the HTML5 Local Storage database, respectively.

Additionally, in Fig. 3 we show how NoTrace is able to make users aware about which personal, sensitive and fingerprinting information they disclose towards third party sites. Specifically, that Figure shows which private information are leaked (i.e. Email address, UserID, Age, Full name and Country), their manner of leakage (i.e., HTTP Referer field, HTTP cookies, Third party cookie and so on), the involved domains and requested URLs.

Moreover, users should be able to decide whether and what actions to take, whether or not Web sites can share information about them with corporate affiliates or third party servers. While other tools in this field only allow to specify blacklist or whitelist, as discussed in Section 4.1, we allowed users to customize their privacy setting. The customization is allowed by default for experienced users. For novice users, instead, when selecting a specific level of protection we let users to exclude from the filtering the techniques that involve a deep impact on the quality of the Web pages returned (i.e., the technique that filter out Web images) and involve page quality break (i.e., the technique that disallows HTTP cookies). We suggest their activation, but the user has the full control on the resultant behaviors. Finally, users have full control on decisions about which Web sites exclude from the filtering, by adding them on-the-fly, to the NoTrace whitelist (through a popup that automatically appears at the top right of the main browser window, as shown in Fig. 2).

4.6. Personalization

An interesting feature of NoTrace is the personalized privacy provided for both novice and expert users. To meet the needs of unexperienced users, instead of asking them to choose among possible actions to preserve privacy, NoTrace automatically detects privacy violations and provides users personalized recommendations about the corresponding countermeasures. Experience users, instead, could be interested in knowing the risks of their navigation and leverage the resultant NoTrace privacy recommendations.

To provide this functionality, we have designed a *Detection* phase, or "*learning period*" in which NoTrace detects all potential privacy violations during Web navigation. Specifically, it studies what sites users they go to and figure out what rules (i.e., techniques) are best applicable for them, ensuring, therefore, the maximal privacy protection at minimal performance cost. We separate the threat detection phase from techniques that could be applied to reduce the potential privacy loss as a user may want to know about what privacy threats are occurring before looking for to do something about it.

To provide personalized recommendations about privacy, NoTrace takes into account three privacy sensitivity categories, named *Low Risk*, *Medium Risk* and *High Risk*, respectively. The goal of this organization is to place threats into a category that will dictate their degree of risk. If a category is assigned a high weight it means that it contains privacy threats that could induce a higher risk on users' navigation.

In general what we want to do is to avoid giving more weight (and consider more dangerous) to items that appear most often during a Web navigation, such as Web images, but that expose users to less risks to their privacy, compared to items that are requested less, but that are widely recognized as highly risky objects (i.e., Web bugs). We have to emphasize, therefore, that to make Web page usable and to avoid page breaks on popular Web sites, we added the threats that expose users to less risks in the Low Risk category, and the threats that involve a serious privacy leakage in the High Risk category (obviously the tradeoff is represented by threats in the Medium Risk category).

Fig. 2. Inspection, through the "Blocked Objects Panel" of third party objects filtered by NoTrace during a Web navigation.

To assign each privacy threat to a specific category, since there is not a taxonomy that states whether a given privacy threat is more dangerous than another one, we take into account the studies that have shown interest and concern, and in general made discussions and suggestions about online privacy. We show each privacy threat and corresponding studies (their references) in Table 2.

We denote with $P_1 \cdots P_n$ n privacy threats and we organize them in three different categories: C_{Low},C_{Medium} and C_{High}, where C_{Low} includes all threats that induce a lower risk on the user's privacy while C_{Medium} and C_{High} include threats with medium and high risk, respectively (this organization reflects our study in Table 2).

Now, if we denote with N_i the number of occurrences of the privacy threats that belongs to the category C_i, with WC_{Low}, WC_{Medium} and WC_{High} the weights that we can assign to the previous defined categories (whose values are set to 20, 30 and 50, respectively), then, the privacy risk (or degree of risk) $PRisk_i$ for the ith category is the following:

$$PRisk_i \doteq \frac{N_i * WC_i}{\sum_j (N_j * WC_j)} * 100$$

where i, $j \in \{Low, Medium, High\}$

In summary, during the Detection phase NoTrace is able to calculate the degree of risk of the navigation of each user by simply looking at his/her activities on the Web. During the (configurable) 2 weeks of learning, NoTrace detects URLs for third party domain servers navigated by users without blocking them, and logs which rules (techniques) are best applicable to calculate the above statistic. Specifically, by weighing the privacy threats, detected during the learning period, it will be able to give to users exactly information about the nature of their navigation and particularly, which percentages of triggered third party activities correspond to activities with low, medium and high risk. If we consider the example described before, Web bugs have been added to the High Risk category, to which we have assigned the highest weight than the other two categories (because it is most dangerous). As result, they will contribute more to the weighted average, than objects (i.e., Web images) belonging to other categories (i.e., Low Risk category).

In Fig. 4 we show to users the calculated percentages. In this way users are informed about the degree of risk of their navigation. Here we can see that the user's training ended informing the user that his/her navigation was composed of 65.28% of objects with high risk, 32.08% of objects with medium risk and 2.64% of objects with low risk. By selecting any of the three provided options, NoTrace will activate the corresponding techniques to reduce the degree of risk of the shown percentages.

We have to emphasize that, as the NoTrace "Full Control" requirement, users can accept these suggestions as well as completely ignore them, deciding by their own and therefore, examining thoroughly the provided techniques (through the NoTrace Control Panel) and activate those that suit their needs.

In summary, in Table 3 we show the comparison between NoTrace and all the analyzed tools, whose details have been discussed in Section 4.1. A $\sim$ symbol in Table means that the technique for the corresponding functionality has not been implemented and that threats may be blocked as consequence of the application of other

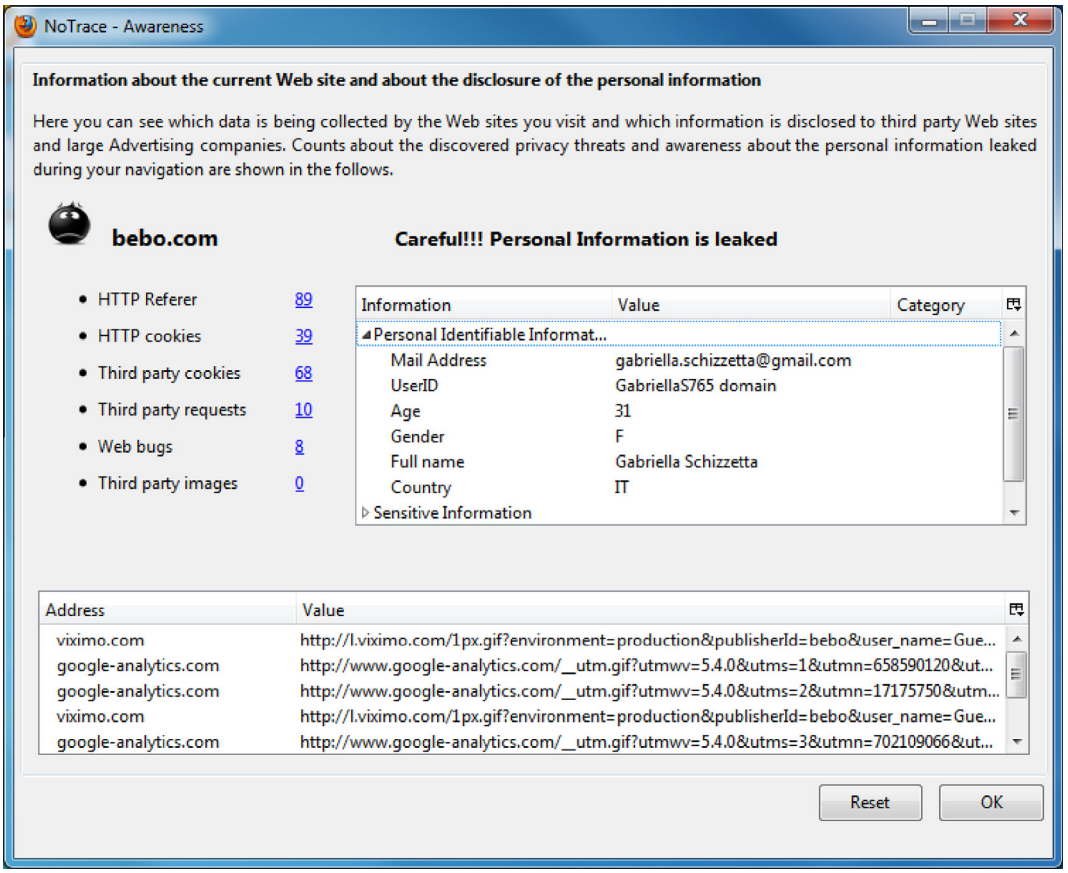


Fig. 3. NoTrace awareness about the leakage of personal, sensitive and fingerprinting information.

Table 2
Classification of the privacy threats into three categories of risk. References represent sources from which we derived the real danger of the corresponding objects. The High Risk category includes threats that refer to the Web Tracking classification, discussed in Section 3.

Degree of Risk	Privacy threats
High Risk	Web bugs and third images [9,50,65,66]
	Third party cookies [9,33,50,65,66]
	Flash cookies [9,38,49,66,73]
	Third party JavaScript executions [9,38,61,65]
	Tracking from *ad*-Networks [9,19,38,65,74]
	Hidden objects and objects from top aggregation servers [38,39]
Medium Risk	Browser Environment Information [33,59–61]
	Personally Identifiable Information (HTTP Cookies, Identifying URLs) [9,42,50,66]
	Personally Identifiable Information (JavaScript executions) [37,38,61]
	Advertisements and Malicious Interfaces [50,69]
Low Risk	HTML5 LocalStorage [51,67]
	Web Images [42]

functionalities (i.e., in general, the technique that filters out all requests for third party domain servers).

This table shows that all tools provide functionalities to block third party requests, while all other techniques to protect privacy of individuals are fully provided by NoTrace, and only partially (when supported), by the other tools. Finally, only NoTrace provides the Content-based filtering mechanism to inspect third party objects in the DOM of the Web pages. Awareness is partially provided by AdBlock Plus, RequestPolicy, Ghostery, and NoScript (awareness about blocked objects or URLs, but no awareness about personal information leakage), while the personalization, that we have studied by measuring the attitudes of users, is fully provided by NoTrace only.

5. Methodology

Our study focuses on three aspects: (1) understanding the nature of privacy leakage, the increasing penetration of third party domains in users' online activities, and the extent of both of them; (2) analyzing the impact of privacy protection techniques on the quality of users' navigation; (3) analyzing the impact of privacy protection techniques on users' perceived experience and performances. We describe the methodology used for the study here while the results are presented in the following Section.

5.1. Data sets

To obtain a realistic data set, we followed the same method as in [42] and selected top Web sites listed by

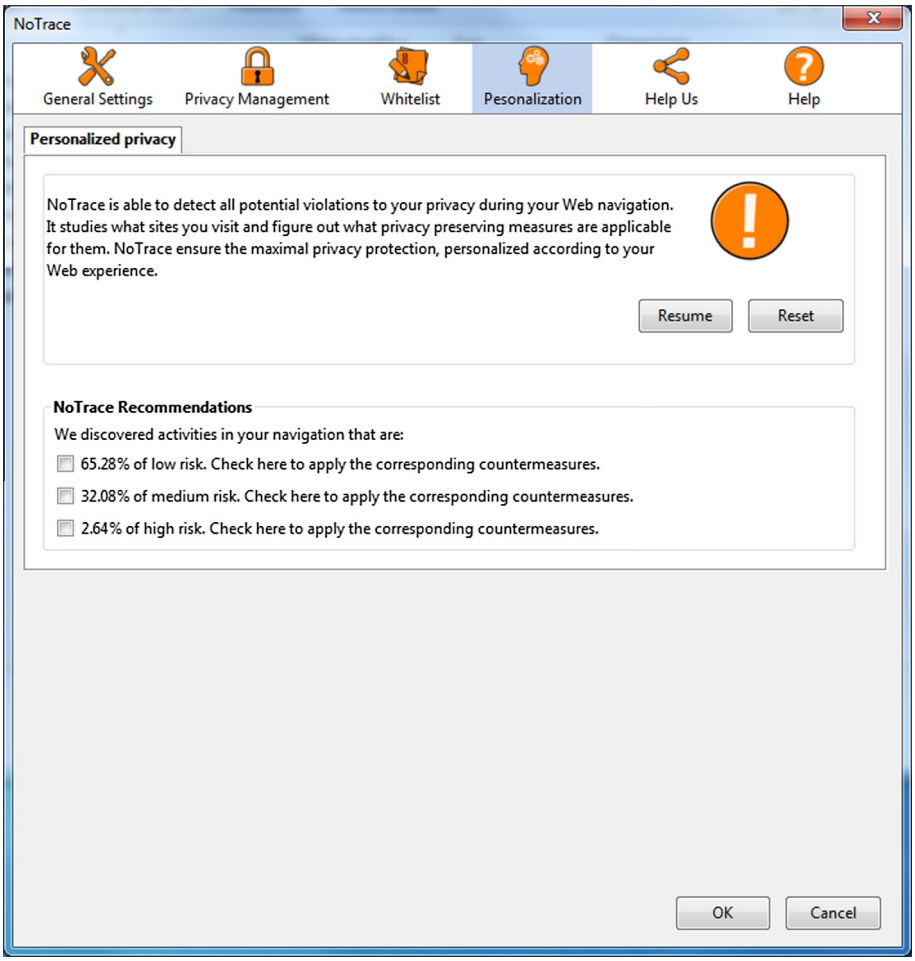


Fig. 4. NoTrace personalized protection: the user exhibits a navigation with the 2.64% of objects at a low risk and with the 65.28% of objects that are highly risky.

Alexa [75]. Specifically, we used two sets of Web sites for this study (See Table 4). The first, named "AllCategories", is a set of English-language sites chosen across all categories from Alexa's popular sites. This set included 1600 pages from each of 16 categories: arts, business, computers, games, health, home, news, recreation, reference, regional, science, shopping, kids & teens, sports, society and world (i.e., Italian Web sites). The second set (we extended the set used in [42,39]), named "Fiduciary", is composed of other 100 Web sites and focusses on pages that involve managing of personal fiduciary information. A high percent of Fiduciary sites [76] suffers from at least one privacy design flaw, by making users vulnerable to off-line attacks that can compromise their personal information, such as social security numbers, account numbers, birthdays, etc.

As our basic measuring methodology to gather information about page downloads, we used the Firefox browser augmented by the "Pagestats" extension [77]. This extension logs information about when each HTTP request was made and the response received and writes it to a log file when all objects for a page are loaded. The interface allows the extension to run the browser in batch mode by specifying a list of sites to access. This extension, used to retrieve 1700 pages of our two data set (built in September/October 2010 and updated in January/February 2011) involved over 140,000 URLs to be analyzed (see Table 4), as discussed in the next Section.

For the object retrievals we have defined four types of retrieval methods according to the tests to be performed: (a) the *Normal* retrieval method allows to retrieve original Web pages without any manipulation, (b) the *Detection* retrieval method allows to get statistics about third party objects without removing/blocking them, (c) the *OnlyText* retrieval method retrieves pages by reducing them to only text (that is, by removing most of the embedded resources), and finally, (d) the *Blocking* retrieval method retrieves Web pages and blocks requests for unwanted content.

5.2. Extent of privacy loss

To analyze the extent of privacy loss and the penetration of third party domains we used Pagestats to retrieve pages contained in both our two data sets (*AllCategories and Fiduciary*) and the *Detection* retrieval method to retrieve them after manipulations of NoTrace techniques.

Table 3
Summary of supported functionalities and system features of privacy tools. The Header Management column refers to the Referer and User Agent header fields.

Tool	Techniques								Requirements		
	Header Mngmnt	HTTP cookies	Flash cookies	Web bugs	Local storage	OPT-OUT cookies	3Rd party requests	Ads full control	Awareness and methods	Filtering	Support and personalization
NoTrace	√	√	√	√	√	√	√	√	√	URL Content-based	√
AdBlock	–	–	–	∼	–	–	√	√	–	URL	–
Ghostery	–	–	√	√	–	√	√	√[a]	√[b]	URL	–
NoScript	–	–	–	∼	–	–	√	√[c]	–	URL	–
RequestPolicy	–	–	–	∼	–	–	√	–	√[d]	URL	–
Adblocker	–	–	–	∼	–	–	√	√	√[4]	URL	–

[a] Ads blocked for a specified set of Advertising companies only.
[b] Awareness about names of the blocked Advertising companies only.
[c] Awareness about counts of blocked scripts only.
[d] Awareness about blocked URLs only.

Table 4
Category breakdown by top level domains: number of URLs selected from Alexa categories and the total number of embedded resources requested by them.

Category	com	org	gov	net	edu	other	Total URLs	Total embedded requests
Arts	87	4	0	6	0	3	100	10,413
Business	95	1	1	2	0	1	100	7491
Computers	87	8	0	5	0	0	100	5297
Fiduciary	95	1	1	1	0	2	100	6336
Games	86	3	0	7	0	4	100	9076
Health	64	13	9	2	2	10	100	7117
Home	87	6	4	2	0	1	100	9225
Kids	87	5	4	1	1	2	100	8928
News	89	3	0	3	0	5	100	13,072
Recreation	91	4	0	3	0	2	100	7172
Reference	51	9	5	0	27	8	100	4523
Regional	68	7	2	3	1	19	100	7767
Science	55	19	12	2	4	8	100	7789
Shopping	94	0	0	0	0	6	100	7933
Society	73	12	8	1	0	6	100	8733
Sport	85	2	0	3	0	10	100	11,020
World	33	3	0	4	0	60	100	8750
Total	1327	100	46	45	35	147	1700	140,390

We have analyzed how privacy can be undermined because of the tracking of users' activities through third party cookies, Flash cookies, advertisements, Web bugs and executions of third JavaScript code by *ad*-networks and Top-10 third party domain servers. We have also analyzed HTTP headers to inspect how much frequently personal identifiable information are being sent to third party domains and aggregators, also through identifying URLs. Statistics about objects triggered by these techniques are saved both in Pagestats and NoTrace log files, for next analysis.

5.3. Impact on page quality

We present here our methodology on the impact of privacy protection techniques on the quality of the Web pages returned. The quality of Web pages is typically a subjective metric, and, to evaluate it, we have designed an objective metric, according to the two following considerations. Intuitively, the worse result in terms of page quality is achieved if all content from pages of our data sets is reduced to nearly mere test (because of several filtering techniques enabled). Being too severe, this degradation serves as a *lower bound* on the quality degradation.

Likewise, the best result in terms of high quality is achieved if all content is downloaded without manipulations (because of no manipulation with NoTrace), serving as an *upper bound* on the quality degradation.

To get an estimation of the upper bound we have used the Normal retrieval method by measuring the byte sessions length of the corresponding requests submitted by Pagestats, without any filtering/manipulation, while to get a measure of the lower bound we have used the Only-Text retrieval method, by enabling in NoTrace all its filtering techniques.

Next, we have measured the quality of Web pages returned when specific techniques are being applied on the stream of HTTP requests from our data sets, that is the *noad*, *nojs*, *noimg*, *noidentURLs* and *noidheader* techniques, and finally, the *notracking* super set of techniques (that includes *no3js, no3cookie, nowebbug, no3hiddenobj, notop*) that allows users to protect their privacy against the behavioral advertising issue (their description has been presented in Table 1). We have excluded the techniques

that access the file system and that do not involve a large number of bytes filtered.

We computed the cumulative distribution functions (CDFs) of the byte session length for all the experiments, both when the Normal and OnlyText retrieval methods were employed as well as when the techniques, mentioned before, were enabled.

The idea is that a CDF of a given technique may exhibit a behavior that is close to the lower bound, stating that the corresponding technique provides protection at a high quality cost, or may exhibit a behavior close to the upper bound stating that the given technique has a minimum impact on page quality.

While we can visually compare distributions by looking at the behaviors of their corresponding curves, results can be quite subjective. The Kolmogorov-Smirnoff test [78] (K-S Test) is a means to mathematically compare two distributions and determine if there is significant difference between them. Specifically, the K-S test compare the maximum probability difference between two CDFs, that is the maximum vertical difference, or D-value, between the analyzed curves. For our analysis we computed two values for each test, the maximum distance from the upper bound or D_U-value (i.e., the statistical difference between the CDF of the byte session length under Normal conditions and the CDF of the tested technique) and the maximum distance from the lower bound or D_L-value (i.e., the statistical difference between the CDF of the byte session length using the OnlyText retrieval method and the CDF of the tested technique).

5.4. Impact on users experience and performance

We carried out a preliminary experiment to test the efficiency of the proposed privacy protection techniques. Specifically, timing measurements are taken at the extension to understand the performance impact of the techniques as the extension could involve a slowdown in handling privacy for a large number of user requests.

Next, we have calculated the gain in terms of response time when third party objects are being removed from users' requests, and to this aim we have defined a metric that can be computed based on the objects retrieved for a page when a given technique is applied relative to the objects retrieved under normal conditions (i.e., by using the Normal retrieval method).

To study the impact on user experience we have used the Pagestats extension to retrieve contents of pages of our data set by using the two following retrieval methods: (1) Normal, without any technique enabled. (2) Blocking, with the following privacy protection techniques enabled: *no3js, no3hiddenobj, noad, nowebbug, no3objnoid, no3img, notop*.

6. Results and analysis

We present here the results of our study, performed through different classes of experiments, and what can be inferred from them. The objective of the first experiment (in Section 6.1) is to analyze the extent of the privacy leakage, providing deep insights about the percentage of third party objects we have discovered (categorized as privacy threats in Section 3) in our data sets and the penetration of third party sites into first party sites. The second (in Section 6.2) aims to show the impact of the provided techniques to protect privacy on the users experience. In the third (in Section 6.3) we evaluate the impact of privacy protection techniques on the quality of the Web pages returned. The fourth set of experiments (in Section 6.4) has the main goal to compare the performance and the effectiveness of our proposed approach against other solutions in the same field. Finally, in Section 6.5 we carried out an experiment to thoroughly test the efficacy of NoTrace on the most popular Web sites with the aim to prove its soundness and completeness.

6.1. Analysis of the extent of the privacy loss

For this category of tests we have organized third party sites in Families according to their acquisition from large business companies, starting as in [38], and enhancing this study with new results, obtained by public and online information as well as by results from our data sets. The greatest concern about these Families is about the wide range of possibilities (a large number of Web sites that they control) through which they can retrieve information about users' behaviors on the Web.

By using the Detection retrieval method (running NoTrace in Detection mode) we have detected the privacy threats shown in Tables 5 and 6. The percentage of the third JavaScript executions threat range from 2.0% for Fiduciary to 3.32%, 3.50%, 3.55% and 3.78% for Society, Arts, News and Home, respectively. The percentage of the hidden scripts by ad-networks ranges from 0.23% for Home to 0.57% for Fiduciary. The percentage of Web bugs ranges from 0.79% for Health to 1.43% for Shopping.

In general, the results show that the most critical categories are Arts, Home, News and Sport, since they contain the largest number of privacy threats (shown in Fig. 5). Moreover, the worst result, in terms of privacy loss is that the largest number of hidden scripts is exhibited by the Fiduciary category, that, as mentioned in the previous Section, contains all the sites for which users enter personal and confidential data. The World (i.e., Italian Web sites) category shows a high number of Flash cookies and, finally, the 33.5% of Web bugs is for the `google-analytics/__utm.gif` image.

Table 8 shows the Top-10 URLs that are retrieved most frequently from the Top-10 hidden nodes (i.e., sites not directly queried by end users, but silently by first party servers) listed in Table 7. Frequencies are calculated on the total number of HTTP requests analyzed (i.e., 140390 requests). An immediate feedback about this experiment is related to the extension of Google, since it holds the first three positions in the top ten list and a fourth among the other ones.

To provide a precise measure of the extent of Google in terms of controlled business/advertising companies, we have organized third party domains in Families and we have studied their incidence in our data sets. The end result, shown in Table 9, is that the Google Family has a reach

of nearly 51% on average amongst the set of all defined Families. The Other column include the sum of statistics about other *ad*-serving companies discovered in our data set, that is Zedo,[16] Quantserve,[17] AudienceScience,[18] comScore Inc.,[19] AddThis[20] and Facebook.[21]

In Fig. 6, we compare statistics about the analyzed Families, by highlighting two important aspects: the high variability of data for Yahoo! and Microsoft, that highly control a small set of pages and, once again, the undisputed domain of the Google Family.

We carried out a last experiment to analyze the depth of the penetration of third party domains in terms of the number of independent domains accessed by first party servers. As shown in Table 10, the most critical categories are Arts, Computers, Games, Home, News, Regional and Sports since they include Web pages (at least one) that inconceivable contain more than 100 requests for third party domain servers. This result is puzzling if we consider that in the News category 80 out of 100 Web sites contain at least 10 requests that access third party servers and that the average number of independent access varies between a minimum of 5.23 (for the Fiduciary category) to a maximum of 23.81 (for the News category).

In summary, our experiments, here, show that a large fraction of user's activities is monitored by large business companies, as established by results in Tables 5–7, and that the penetration of third party domains first studied in [38], continue to increase at a fast pace, with the dominant role in the tracking market played by Google.

6.2. Impact on users experience

The impact that the application of our techniques would have on user experienced download time is the main goal of the following study.

For this class of experiments we have taken into account the most critic categories as established in the previous experiment and we have performed a comparison among the three different retrieval methods described in Section 5.1.

Results in Table 11 show that third party objects also play a significant role in increasing the number of objects per page, as well as download times. As we can see from this table, by enabling NoTrace, we are able to block third party objects and, at the same time, we can save from the 42.75% (for the Business category) to 57.86% (for the Arts category) of the total Megabyte transferred in downloading Web pages.

Table 11 shows the minimum impact of NoTrace on the user perceived latency. In fact, as we can see from that table, when NoTrace is in Detection mode the download time undergoes an increase of a few hundred milliseconds for the Business category up to a maximum of 2 s for the Sport category, with an overall average of 1.3 s compared to the *Normal* navigation. We have also to emphasize that the personalization comes at a cost, and therefore, with further tests, we get that the overall overhead is about 3 s (in the worst case). The reason is that the Detection time incorporates the times of different techniques for which statistics have to be stored for every HTTP request. It must be emphasized, however, that this further degradation occurs *only* during the Detection phase (in which different statistics have to be calculated) which, on the other hand, is activated only periodically, or according to a user request.

6.3. Impact on page quality

The impact on the page quality of our approach for privacy protection is the main goal of this set of experiments.

As described in Section 5.3, we have computed the CDFs of the byte session length for six privacy protection techniques and we have compared them with the CDFs of the byte session length when the Normal retrieval method is used (we call it the Normal CDF) and when the OnlyText retrieval method is used (we call it the OnlyText CDF).

Fig. 7 shows the behavior of the NoTracking CDF (when the NoTrace *notracking* technique is applied on the stream of HTTP requests) against the behaviors of the Normal and OnlyText CDFs.

By simply looking at the curves, it is not immediately clear whether NoTracking CDF is closer to the OnlyText or to Normal CDF. A more refined quantitative analysis shows, indeed, that the distance between the Normal CDF and the NoTracking CDF is smaller than the distance between the latter and the OnlyText CDF. When comparing *D*-values, shown in Table 12, the end result is that the value of the NoTracking-Normal combination, 0.337, is smaller than the NoTracking-OnlyText combination, 0.524, stating that the NoTracking CDF is a closer match to the Normal than the OnlyText CDF, by finally confirming what claimed before, by simply looking at the graphs.

Results show that, not surprising, the best technique in terms of quality preservation is the *noidheader* technique since its D_V value is the smallest against all other results shown in the first column of the Table 12. Conversely, the worse result is exhibited by the *noimg* technique since its D_V value is the highest against all other results. On the other hand, the overall result about the NoTrace's impact on the quality of Web pages is that all analyzed techniques have a minimum impact since their closer match to the D_U values than the D_L values.

6.4. Performance and effectiveness

To analyze the performances we first performed an initial test to measure the processing overhead of the privacy protection techniques implemented in NoTrace. For this test timing measurements are taken at the NoTrace extension to understand the performance impact of any technique. Table 13 shows that the processing overhead is minimal, often in few milliseconds range, and with half of services incurring an overhead of around half a millisecond. The techniques that took longer are, obviously, those that require the parsing of HTTP responses (i.e., removing advertisements).

[16] http://www.zedo.com/.
[17] http://www.internetanalytic.com/quantserve.com.
[18] http://www.audiencescience.com/.
[19] http://www.comscore.com/.
[20] http://www.addthis.com/.
[21] http://connect.facebook.net/.

Table 5
User behavioral tracking privacy threat presence (%) by categories of the *AllCategories* and fiduciary data sets. The degree of risk reflects the risk taxonomy in Table 2.

Category	Cookie management		Web tracking				
	Flash cookie	Third cookie	Third JS exec.	Top-10 third server	Web bugs	Third images	Hidden scripts by ad-networks
Degree of risk	*High risk*		*High risk*				
Arts	0.41	1.52	3.50	0.21	0.85	6.76	0.30
Business	0.33	1.73	2.65	0.18	1.17	4.12	0.33
Computers	0.24	1.12	2.53	0.12	0.88	7.93	0.23
Fiduciary	0.31	1.17	2.00	0.14	0.68	2.98	0.57
Games	0.33	1.14	2.33	0.16	1.04	4.52	0.29
Health	0.36	1.19	2.63	0.25	0.79	3.59	0.39
Home	0.35	1.71	3.78	0.22	0.99	6.77	0.23
Kid and Teen	0.36	1.33	2.76	0.16	0.85	4.05	0.30
News	0.41	1.65	3.55	0.20	0.96	5.14	0.31
Recreation	0.27	2.04	2.85	0.16	1.32	5.02	0.37
Reference	0.24	1.15	1.87	0.14	1.03	4.02	0.39
Regional	0.38	1.50	2.49	0.14	0.63	3.74	0.42
Science	0.38	1.41	2.39	0.18	0.82	3.14	0.37
Shopping	0.23	1.59	1.84	0.16	1.43	4.35	0.45
Society	0.36	1.45	3.32	0.20	0.85	4.50	0.24
Sports	0.48	1.38	3.05	0.15	0.85	6.03	0.30
World	0.48	1.04	2.16	0.17	1.04	4.49	0.35

Table 6
Browser environment and personal information privacy threats presence (%) by categories of the *AllCategories* and fiduciary data sets. The degree of risk reflects the risk taxonomy in Table 2.

	Third activities and unwanted ads		Personal Information		
	Identifying URLs	Ads	Cookie	Referer	User agent
Degree of risk	*Medium risk*		*Medium risk*		
Arts	5.70	1.48	15.74	29.40	32.26
Business	5.23	1.67	16.36	31.10	33.52
Computers	5.85	1.00	15.95	30.53	32.53
Fiduciary	2.99	1.11	21.20	31.69	33.94
Games	4.34	1.57	19.01	30.98	32.87
Health	4.33	1.17	17.95	31.64	33.50
Home	5.81	1.95	15.55	29.64	31.04
Kid and Teen	5.12	1.33	19.63	30.19	32.50
News	5.96	1.94	15.01	30.72	32.34
Recreation	5.01	1.47	18.10	30.24	31.66
Reference	3.88	1.04	15.07	33.96	35.99
Regional	3.66	1.43	16.52	33.12	34.73
Science	3.98	1.16	18.57	32.10	33.87
Shopping	3.34	0.88	19.61	31.49	33.16
Society	5.63	1.49	16.80	30.77	32.64
Sports	5.66	2.24	15.60	30.30	32.64
World	4.48	1.15	16.16	32.59	34.42

In the second test we have compared the effectiveness of NoTrace with other tools in the same field. Specifically, we have compared NoTrace with Adblock Plus and Adblocker [79] for effectiveness about the advertisement filtering process, with NoScript for effectiveness about JavaScript code execution blocking, with Ghostery [48] and RequestPolicy [47] for effectiveness about the behavioral advertising. To compare NoTrace with Adblock Plus we used its rules to filter out advertisements from the Web pages of our data set. Specifically, we downloaded the Fanboy's List[22] (English) suggested as a suitable subscription for English language.

Before commenting the Table 14, we have to remember that the *notracking* technique of NoTrace, refers to the third party activities we addressed with the following set of techniques: *no3js, nowebbug, no3cookie, no3hiddenobj, notop*. In addition, by also taking into account the *noad* technique we obtain a superset of all the other filters applied by the competitors, where no additional filter/technique can be added.

As we can see from Table 14 turning off third party objects also improve performance as the lower number and bytes downloaded per page. Specifically, Adblock Plus cuts the number of objects served by about the 22.1%, NoScript is able to cut the 43.71%, Ghostery the 25.85%, Adblocker the 13.22%, RequestPolicy the 56.02% and, finally, NoTrace the 47.84%. The worst result is for Adblocker, with cut of only 13.22%.

[22] http://adblockplus.org/en/subscriptions.

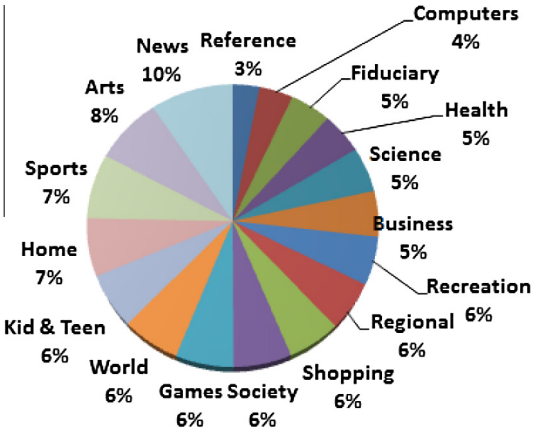


Fig. 5. Category breakdown by privacy threat frequencies. The most critical categories: Arts, home, news and sports.

Table 7
Top-10 hidden third party sites and their frequency (%) in our data sets.

Family	Hidden third sites	Frequency (%)
Google	doubleclick.net	3.01
Google	google-analytics.com	1.53
Google	2mdn.net	1.10
Yahoo!	yimg.com	1.01
Comscore Beacon	scorecardresearch.com	0.96
Google	googlesyndication.com	0.94
Quantcast	quantserve.com	0.74
Yahoo!	yieldmanager.com	0.70
Microsoft	atdmt.com	0.51
RevenueScience	revsci.net	0.42

Table 8
Top-10 third party JavaScript code and their frequency (%) in our data sets.

Hidden third party JavaScript code	Frequency (%)
google-analytics.com/ga.js	0.55
scorecardresearch.com/beacon.js	0.36
quantserve.com/quant.js	0.26
swfobject.js [Google]	0.19
pagead2.googlesyndication.com/pagead/show_ads.js	0.18
doubleclick.net/pagead/test_domain.js	0.18
2mdn.net/879366/flashwrite_1_2.js	0.17
s_code.js [Omniture]	0.16
revsci.net/gateway/gw.js	0.12
google-analytics.com/urchin.js	0.11

Our experiments here show that NoScript and RequestPolicy are able to remove the largest fraction of third party objects, but it is worth to say that NoScript blocks, regardless of the real danger of detected objects, all JavaScript code (by allowing URLs being whitelisted when the quality and the functionality of the Web page break), while RequestPolicy has a stricter set of rules. For example, it blocks all requests that contain /*media*/ in the corresponding URLs, by avoiding the page break for most of the popular Web pages only because of they are included, by default, in the startup whitelist. Conversely, all other analyzed extensions start with a blank whitelist. Anyway, the difference with NoTrace, in terms of size of blocked objects, is only 2%. On the other hand, when compared to Adblock, Ghostery, and Adblocker, NoTrace is able to block the larger number of third party objects.

In summary, our tool is able to remove "intelligently" the largest fraction of dangerous objects, showing better results when compared with NoScript (that blindly removes all the scripts, even those that are essential to the correct behavior of the page), and showing a slight difference, in terms of blocked objects, when compared with RequestPolicy. Moreover, neither NoScript nor RequestPolicy provide adequate feedback about blocked objects and their real dangerousness.

As discussed in Section 2, nor of the analyzed extensions is able to fully face all known privacy concerns. A more secure navigation would require the installation and configuration of multiple of them (sometimes conflicting) into the browser. A drawback, anticipated in Section 3, is about the degradation of performance of Firefox when it loads and deals with multiple extensions. To prove that, we have performed an experiment that aims to compare the performance of Firefox when four extensions were loaded and enabled (AdBlock Plus, NoScript, Ghostery and RequestPolicy) with specific techniques (i.e.,advertisements and Web bugs filtering, JavaScript code execution blocking and third party JavaScript execution by *ad*-networks blocking) against its performance when only NoTrace is loaded, providing "all in one" the same functionalities.

The end result is that the test duration time was of 17 s (to request 13655 URLs and to filter them) for the multiple-installation against only 7 s for NoTrace, showing an evident gain in terms of system responsiveness and better users' experiences.

We would like also to add that NoTrace efficiency is not traded off with modularity. In fact, NoTrace leverages the Cross Platform Component Object Model (XPCOM) framework,[23] that allows the development of modular software and provides tools to create, assemble and manipulate components at run-time. Specifically, we have implemented three different components: the first implements techniques that manages HTTP requests/responses headers, the second component explicitly manages the Content-based filtering mechanism, applying on-the-fly transformations before the browser rendering begins, while the third manages the URL-based blocking filtering mechanism. The integration of new countermeasures can be realized by implementing the new countermeasure as part of one of the three NoTrace' components (chosen according to the kind of resources to manipulate) by only developing the JavaScript functions that represent the logic of the new functionality to offer to protect privacy. Otherwise, whether new types of mechanisms are required, it is possible to design and implement a new XPCOM component, that has to be registered into Mozilla, and has to implement the needed interfaces and the corresponding methods (i.e., nsIObserver and nsISupports).

[23] https://developer.mozilla.org/en/XPCOM.

Table 9
Families extent (%) for all categories of our data sets. The Other column includes statistics about Zedo, Quantserve, AudienceScience, comScore Inc., AddThis, and Facebook.

Families	Google	Yahoo!	AOL	Omniture	ValueClick	Microsoft	Other
Arts	44.11	14.19	12.54	2.51	0.59	10.92	15.13
Business	51.04	12.71	4.27	7.40	4.93	3.70	15.94
Computers	37.29	10.22	2.34	3.99	0.68	30.57	14.90
Fiduciary	57.09	9.09	2.00	14.36	0.91	5.27	11.27
Games	48.26	12.19	12.81	2.44	0.35	6.96	16.99
Health	64.57	4.38	6.12	2.74	0.91	3.29	17.99
Home	47.94	6.73	16.90	2.15	0.68	8.76	16.85
Kid and Teen	59.65	4.54	3.41	4.72	0.35	4.72	22.62
News	40.48	14.52	6.52	3.40	1.67	12.35	21.07
Recreation	44.02	10.28	11.15	4.08	0.73	3.35	26.38
Reference	49.59	21.07	6.97	2.27	0.49	1.46	18.15
Regional	39.13	13.33	8.40	6.45	1.10	16.30	15.28
Science	61.33	5.96	1.56	4.79	0.39	1.46	24.51
Shopping	51.67	12.68	6.81	13.48	0.00	3.74	11.62
Society	60.21	10.95	7.23	5.29	1.94	2.46	11.92
Sports	37.30	32.71	12.05	1.83	0.85	4.37	10.89
World	79.19	8.67	0.35	3.70	0.92	3.35	3.82
Mean	51.35	12.01	7.14	5.04	1.03	7.24	16.20

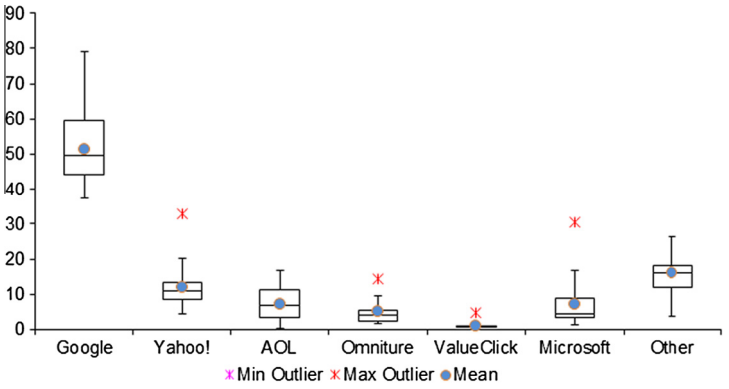


Fig. 6. Statistics about the most famous families from our data sets. The domination of Google is overpowering.

Table 10
Depth of third party penetration amongst first party servers contained in our data sets.

	Number of accessed third party domains				
	⩾10	⩾20	⩾30	⩾50	⩾100
Arts	67	36	19	7	3
Business	29	21	12	4	0
Computers	22	13	9	5	1
Fiduciary	18	5	2	0	0
Games	36	20	10	6	2
Health	37	17	8	2	0
Home	56	30	17	4	1
Kid & Teen	42	15	8	1	0
News	80	63	32	9	1
Recreation	37	21	13	3	0
Reference	19	11	3	2	0
Regional	32	19	9	3	1
Science	32	18	10	1	0
Shopping	19	10	3	1	0
Society	45	29	14	1	0
Sports	60	35	21	10	1
World	27	15	3	0	0

6.5. Effectiveness on the most popular Web sites

When evaluating the number and the type of vulnerabilities found in each category of sites in the experiment described in Section 6.1 we ignored the possibility of false positive and false negative detection by NoTrace. We drawn out useful numbers, but to be more accurate we performed the following further experiment.

We have "manually" analyzed the first 25 most popular Web sites listed by Alexa[24] (with nearly 2000 embedded resources) from which we excluded the Google Web site and all its nationalized versions.

By using Pagestats in a batch mode, we issued requests for these sites enabling in NoTrace the techniques to address the behavioral advertising (i.e., *no3js*, *no3cookie*, *nowebbug*, *no3hiddenobj*, *notop*), referred as *notracking* technique in Section 5.3, and to block advertisements (i.e., *noad*) and cookies (i.e., *nocookie*). We have also inspected if disabling (first and third party) cookies involves a break of visited Web sites in terms of page functioning.

[24] Data retrieved on 20th October, 2011.

Table 11
The impact of third party objects on the Web page download time when NoTrace is enabled.

Category	Retrieval method	Median download time (ms)	Object savings	
			Size (MB)	Variation (%)
Arts	Normal	4347	110.12	−57.86
	Detection	5381	–	
	Blocking	2953	46.41	
Business	Normal	4329	57.61	−42.75
	Detection	5016		
	Blocking	3016	32.98	
Home	Normal	4643	65.18	−52.66
	Detection	5808		
	Blocking	2838	30.85	
News	Normal	5228	92.11	−45.55
	Detection	6972	–	
	Blocking	4028	50.15	
Sport	Normal	3918.5	123.82	−57.15
	Detection	5914	–	
	Blocking	2826	53.05	

Before describing the experiment and its results, an important consideration is about three different types of objects retrieved during the experiment itself: firstly, small scripts or simple JavaScript libraries used by pages to control their UI and whose filtering could involve their rendering almost useless; secondly, the identifying URLs, which could be any kind of safe query parameter and not requests that transmit personally identifiable information; finally, third party images, which could be load balancing servers and whose filtering could significantly deteriorate the quality of Web pages returned.

Therefore, to offer a meaningful conclusion about NoTrace's effect on privacy we have calculated: (a) the percentage of objects blocked by the *notracking* and *noad* techniques ("Standard" Blocking column in Table 15); (b) the percentage of objects blocked by the *notracking*, *noad* and, in addition, by the techniques that block identifying URLs and third party images ("Overall" Blocking column in Table 15). Finally, to make a separation between needed objects for a high quality and functional rendering of a Web page and third party objects, we have inspected all domains (of the 25 Web sites analyzed) to be whitelisted (given their role of CDNs) to avoid both quality and functionality break. After the creation of this whitelist we have calculated the percentage of objects blocked by all the previous techniques ("Intelligent" Blocking column in Table 15) by excluding from the filtering all objects served by the domains included in the whitelist.

Table 12
Measure of the Page Quality impact. The D_U value represents the statistical result for the upper bound on the page quality while the D_L the statistical result for the lower bound.

Technique	Description	Normal (D_U value)	OnlyText (D_L value)
notracking	Behavioral advertising	0.337	0.524
noad	Removing adverts	0.100	0.664
nojs	JavaScript executions	0.197	0.660
noimg	Removing images	0.370	0.436
noIdentURLs	Removing personal information	0.232	0.627
noidheader	Removing personal information	0.045	0.674

As an example, if we look at the Table 15, we can see that the contents served by the `taobao.com` Web site also come from three different third party entities (i.e., CDNs, serving mostly Web images). If we look at the percentages we can see that the "Standard" and the "Overall" values are 20.3% and 98.9%, respectively, showing that the quality of the page could be highly degraded even without any important improvement on privacy. In fact, in this example we could wrongly remove the 76.3% of the triggered objects whereas only the 2.3% is really dangerous. Results of this experiment can be summarized as follows:

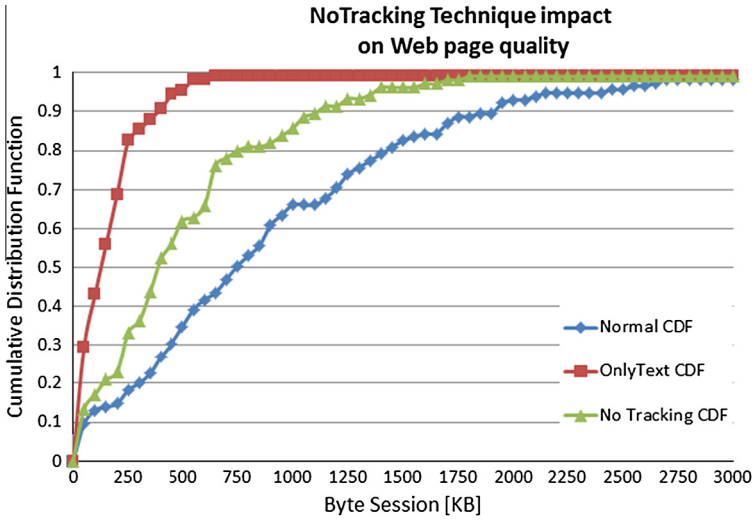


Fig. 7. Cumulative distribution functions of the byte session length for the *notracking* superset of privacy protection techniques.

Table 13
Overhead observed for third party activities (time in ms).

Category	Third cookies	Third JS code	Ident. URLs	Adverts	Hidden third servers	Top-10 servers	Web bugs	Pers. info
Arts	0.055	0.993	0.032	4.336	0.244	0.043	0.109	0.058
Business	0.027	0.942	0.023	4.498	0.250	0.041	0.122	0.063
Computers	0.042	0.934	0.039	4.445	0.200	0.033	0.075	0.053
Fiduciary	0.029	0.956	0.061	4.559	0.229	0.116	0.088	0.054
Games	0.073	0.965	0.036	4.152	0.211	0.057	0.103	0.048
Health	0.055	1.043	0.032	3.720	0.247	0.035	0.088	0.056
Home	0.032	0.965	0.027	3.554	0.259	0.057	0.145	0.060
Kids and Teens	0.031	0.984	0.030	5.165	0.206	0.057	0.107	0.055
News	0.053	0.718	0.031	3.670	0.261	0.056	0.111	0.064
Recreation	0.059	1.002	0.028	5.309	0.338	0.046	0.092	0.061
Reference	0.044	0.919	0.050	4.903	0.326	0.037	0.131	0.055
Regional	0.033	0.987	0.028	4.240	0.226	0.066	0.117	0.136
Science	0.039	1.040	0.030	5.168	0.213	0.055	0.116	0.062
Shopping	0.040	0.008	0.022	4.877	0.263	0.044	0.085	0.067
Society	0.023	0.984	0.036	4.596	0.211	0.053	0.135	0.057
Sports	0.038	0.958	0.057	4.572	0.273	0.042	0.117	0.132
World	0.052	0.945	0.035	4.119	0.177	0.067	0.095	0.052
Mean	0.042	0.903	0.035	4.464	0.243	0.053	0.108	0.067
St. Dev	0.014	0.241	0.011	0.524	0.043	0.019	0.019	0.026

Table 14
Comparison with Adblock Plus, Adblocker, NoScript and Ghostery in filtering third party objects from the News category of the *AllCategories* data set.

Tool	Techniques enabled in each tool	Objects	
		Count	Size (MB)
Normal	Retrieval under normal conditions, no tool installed	13,372	108.31
AdBlock Plus	Filtering advertisements	10,309	75.86
NoScript	Blocking scripts	7527	50.89
Ghostery	Disabling Web bugs and third party tracking	9915	75.57
NoTrace	Disabling Web bugs, third party tracking, blocking scripts, filtering advertisements	6906	46.87
Adblocker	Filtering advertisements	11,604	84.07
RequestPolicy	Disabling third party tracking	6729	49.05

1. Allowing both third party images and identifying URLs and blocking all other third party objects involves a low impact on the quality of Web pages as well as an increased exposition to privacy risks.
2. Enabling all techniques for privacy protection involves a serious impact on the quality of Web pages but a strong defense against privacy risks.
3. Allowing third images and identifying URLs only from domains included in the whitelist ensures a correct tradeoff between quality and privacy preservation.

It must be noted that it is possible to further improve the quality of the Web pages by also allowing third party scripts whose corresponding domains are being included in the whitelist. The drawback here is about possible false negatives that this action can involve.

With regard to the presence of false positives and negatives, the result is that their incidence is very low, since we have obtained a percentage of false positives and negatives of 1.17% and 0.37%, respectively. The majority of false positives detected by NoTrace is for URLs whose domains have been included in the whitelist. In fact, as you can see in Table 15, looking at the false positives of the `facebook.com` Web site, the meaning of 55/0 is that we can fully delete all detected 55 false positives by simply adding the sites specified in the second column (i.e., `fbcdn.net` and `akamaihd.net`) to the whitelist.

The remaining number of false positives is for first party (and therefore theoretically not risky) identifying URLs, that sometimes refer to requests for Web bugs. A feasible solution to eliminate these "errors" is to envision a technique that does not brutally remove URLs with character code such as '?', '=', or '&', but that is able to remove only URLs that involve leakage of userids, usernames, and piece of private information to third party servers [33].

The false negatives inspected in our experiment are related to two types of requests: (a) "first party" requests for third party objects and (b) "third party" requests that hide third party activities (i.e., Web bugs) or requests for objects for which we do not have enabled any technique (requests for CSS, swf, JSON files, and so on). An example of request of the first type (from the `qq.com` Web site) tries to inspect the IP address of the client making such a request (i.e., http://fw.qq.com/ipaddress), while an example of a request of the second type (from the `mail.ru` Web site) tries to transparently request a Web bug (i.e., http://www.tns-counter.ru/V13a****mail_ru/ru/CP1251/tmsec = mail_main/). By enabling the *no3obj* technique (shown in Table 1), implemented in NoTrace but not tested here, we could drastically reduce the number of these false negatives.

Finally, it is likely that the Web site's behavior is dependent on cookies that are unexpectedly being rejected or removed. In the seventh column of Table 15 we show which sites do not allow access to users when both first and third party cookies are being forbidden (i.e., 1 means no access).

Table 15
Analysis of NoTrace effectiveness on the Alexa's top 25 sites.

Web site	Whitelist	Blocked objects (%)			Cookies[a]	False pos.	False neg.
		Standard	Overall	Intelligent	3rd/1st	yes whitelist/no whitelist	
facebook.com	fbcdn.net akamaihd.net	1.0	54.5	4.0	0/1	55/0	0
youtube.com	ytimg.com	43.2	56.8	54.0	0/0	6/0	0
yahoo.com	yimg.com	30.2	71.7	35.9	0/0	29/0	0
baidu.com	–	28.6	28.6	28.6	0/0	3/0	0
wikipedia.org	wikimedia.org	0.0	93.3	0.0	0/0	13/0[b]	0
blogspot.com	blogger.com google.com	12.8	18.0	12.8	0/0	5/0	0
live.com	wlxrs.com msn.com hotmail.com	46.4	90.5	53.6	1/ 1	55/0	0
twitter.com	twimg.com	35.0	75.0	35.0	1/1	9/0	0
qq.com	gtimg.com	15.1	47.3	15.1	0/0	30/0	2
msn.com	s-msn.com	17.8	23.9	17.8	0/0	3/2	0
yahoo.cp.jp	yimg.jp	20.0	88.6	25.7	0/0	26/0	2
linkedin.com	–	33.3	37.7	34.8	1/1	7/1	0
taobao.com	tbcdn.cn taobaocdn.com mmcdn.cn	20.3	98.9	22.6	0/0	134/0	0
sina.com	sinaimg.cn	30.9	90.1	37.5	0/0	73/2	1
wordpress.com	wp.com	26.2	95.2	31.0	0/0	27/0	0
ebay.com	ebaystatic.com ebayimg.com	17.5	94.7	21.1	0/0	50/0	0
yandex.ru	yandex.st yandex.net	28.2	82.1	28.2	0/0	19/2	0
163.com	netease.com	32.0	96.0	41.6	0/0	121/2	0
microsoft.com	–	22.8	26.3	26.3	0/0	2/2	0
weibo.com	sinajs.cn sinaimg.cn	8.5	93.0	9.9	0/0	59/0	0
mail.ru	imgsmail.ru	9.8	45.9	9.8	0/0	42/0	2
flickr.com	yimg.com	61.3	98.8	66.3	0/0	20/0	0
apple.com	–	34.2	34.2	34.2	0/0	0/0	0
imdb.com	media-imdb.com	29.3	79.3	40.6	0/0	51/10	0
bbc.co.uk	bbcimg.co.uk static.bbc.co.uk	44.4	81.1	54.4	0/0	24/0	0

[a] Means break.
[b] This example shows how the *no3img* technique could uselessly break the quality of a page as all detected objects are third party images (whose domain is in the whitelist).

7. Related work

Several systems and tools that address the problem of online privacy exist in literature. Specifically, several techniques [42] and tools, both intermediary-based [43] and client-based [45–47,55,48,53,**?**,58], have been designed and implemented. We do not discuss the client-side solutions as their detailed description presented in Section 4.1.

Most of the existing functionalities implemented directly inside browsers to provide private browsing aim to reduce the amount of information that can be saved on users' computers, that is local cache, histories, credentials and so on. Examples include Incognito by Google Chrome,[25] InPrivate Browsing by Internet Explore 9,[26] Private Browsing by Mozilla Firefox.[27] In a private browsing session, browsers will not keep any browser history, search history, download history, Web form history, cookies, or temporary Internet files.

The simple user-driven mechanism of privacy protection provided by the Network Advertising Initiative shows limitations in terms of users protection from tracking. Users do not widely know or use the opt-out mechanism. The main reason is that this mechanism does not persist reliably with the counterintuitive effect that by deleting the standard cookies (privacy threats) will also erase the opt-out cookies, with the net effect of requiring further actions, that are felt as duplicated by the unexperienced user. Moreover, because there is no generalized mechanism, and an opt-out cookie is required for any advertising company, the process appears as a tedious task making users reluctant to use it. To overcome this limitation and to preserve preferences for online behavioral advertising, the NAI Consumer Opt Out Protector Firefox add-on [80] has been implemented for all participating NAI member companies. A drawback is that users must first set NAI opt-out preferences (either before or after installing the add on) through the NAI opt-out tool to allow the corresponding extension to deal with.

In [74] authors present an approach to help users to understand what third party companies know about users personal information. They use a JavaScript code embed-

[25] http://www.google.com/support/chrome/bin/answer.py?answer=95464.
[26] http://windows.microsoft.com/en-US/internet-explorer/products/i.e.-9/features/in-private.
[27] http://support.mozilla.com/en-US/kb/PrivateBrowsing.

ded in the Web page a user visit. This script accesses to the user's Web history and analyzes sites to discover third party sites. It is a server-side mechanism that simply makes users aware of privacy risks without providing instruments to face them.

Finally, the Mozilla Firefox browser exhibits a new feature that allows to tackle the behavioral advertising issue. Setting up a browser preference, users manifest their desire to opt-out of third party advertising-based tracking by transmitting a Do-Not-Track HTTP header with every click or page view in Firefox. This mechanism shows some drawbacks. First, it requires cooperation of *ad*-networks, that in fact, can ignore the header and track users anyway. This means that users must trust that *ad*-networks (when they effectively will know and support this mechanism) will honor the header and really stop to track users behaviors on the Web. Second, it lacks of awareness and control, since no feedback is provided to users about which *ad*-networks really stop the tracking. In addition users are not allowed to decide to opt-out only for a specific family (Google, Yahoo!, etc.) or for a set of advertising networks instead of for all business/advertising companies, involving disappointment by people that like targeted personalized advertisements [71].

8. Conclusion

In this paper we have presented a background study about which threats could undermine the privacy of individuals during online activities and a measurement-based study that allowed us to provide statistics about the penetration of third party domain servers in the most popular Web sites (represented by our data sets), and about the extent of the Web information leakage (as established by results in Section 6.1).

To carry out the experiments we have used NoTrace, a Firefox extension, whose provided approach for privacy has been described in Section 4. Conversely to existing systems in this field, that only address specific privacy threats, locking users into a kind of walled garden and allowing them only few actions to take, our comprehensive approach is able to detect risky activities on the Web making users aware of which personal information is being gathered, when it happens and in which extent, by suggesting them personalized actions to take. Awareness and full control, in fact, have been recognized as two fundamental requirements when designing tools for privacy protection on the Web [81], as also suggested by the recommendations of FTC, PRC and other important privacy watchdog organizations.

We have also studied the impact of NoTrace on the user perceived download time (Section 6.2) and the impact on the quality of Web pages returned when several filtering techniques were applied on the HTTP stream of requests (Section 6.3), with acceptable results in both experiments. In addition, our measurement study allowed us to prove the efficiency of the implemented techniques and their efficacy through a comparison with the most popular tools in the same field. The result, here, is that NoTrace is able to block the largest number of objects that *represent a threat for the users privacy* (Sections 6.4 and 6.5).

We plan to deeply study emerging privacy threats, that will likely increase in the future, i.e., the HTML5's LocalStorage privacy threat and other privacy concerns related to the use of HTML5 (i.e., misuse of geolocation API, microphone API, and camera API, etc.). We also want to further investigate the potential dangerous activities perpetrated, against unaware users, by third party sites, as well as by first party sites, that in fact, as recently recognized in [33], are not doing a good job in ensuring that private information are not disclosed to other entities (third party sites, data aggregators, affiliates, etc.).

Finally, we have performed an usability study to measure users' attitudes and experiences when using NoTrace and the satisfaction perceived by using its interface. The significant results that we obtained, thanks to the useful suggestions of participants in our study, allowed us to make several changes to the NoTrace interface and to design the enhancements of the existing functionalities to make our tool for privacy protection more effective and usable by a large audience of Internet users.

Acknowledgements

The authors thank the anonymous reviewers who provided valuable, thoughtful and insightful comments that helped to improve the quality and the effectiveness of our initial manuscript.

References

[1] D. Malandrino, V. Scarano, Supportive comprehensive and improved privacy protection for web browsing, in: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT), 2011, pp. 1173–1176.
[2] S.D. Warren, L.D. Brandeis, The right to privacy, Harward Law Review 4 (5) (1890) 193–220.
[3] A. Westin, Privacy and Freedom, New Jork Atheneum, New York, 1967.
[4] R. Leathern, Jupiter Research 2002. Security and Privacy Data, 2002. <http://www.ftc.gov/bcp/workshops/security/020520leathern.pdf>.
[5] Harris, Harris Poll: Privacy and American Business, June 2004.
[6] C. Paine, U.-D. Reips, S. Stieger, A. Joinson, T. Buchanan, 'Internet users' perceptions of 'privacy concerns' and 'privacy actions', International Journal of Human–Computer Studies 65 (6) (2007) 526–536, http://dx.doi.org/10.1016/j.ijhcs.2006.12.001.
[7] UPI, UPI Poll: Concern on Health Privacy, 2007. <http://www.upi.com/Topuist_News/2007/02/21/UPI-Poll-Concern-on-health-privacy/UPI-39291172098800/>.
[8] J. Gomez, T. Pinnick, S. Ashkan, UC Berkeley, School of Information, June 2009. <http://www.knowprivacy.org/report/KnowPrivacy_Final_Report.pdf>.
[9] P.R. Clearinghouse, Fact Sheet 18: Online Privacy: Using the Internet Safely, 2010. <http://www.privacyrights.org/fs/fs18-cyb.htm>.
[10] A.N. Joinson, U.-D. Reips, T. Buchanan, C.B.P. Schofield, Privacy, trust, and self-disclosure online, Human–Computer Interaction 25 (1) (2010) 1–24.
[11] A. Acquisti, J. Grossklags, Privacy and rationality in individual decision making, Security Privacy, IEEE 3 (1) (2005) 26–33.
[12] T. Buchanan, C. Paine, A.N. Joinson, U.-D. Reips, Development of measures of online privacy concern and protection for use on the internet, Journal of the American Society for Information Science and Technology 58 (2007) 157–165.
[13] S. Spiekermann, J. Grossklags, B. Berendt, E-privacy in 2nd generation E-commerce: privacy preferences versus actual behavior, in: ACM Conference on Electronic Commerce, ACM, 2001, pp. 38–47.
[14] R.K. Chellappa, R.G. Sin, Personalization versus privacy: an empirical examination of the online consumer's dilemma, Information Technology and Management 6 (2005) 181–202.

[15] C. Jensen, C. Potts, C. Jensen, Privacy practices of Internet users: self-reports versus observed behavior, International Journal of Human–Computer Studies 63 (1-2) (2005) 203–227.
[16] B. Berendt, O. Günther, S. Spiekermann, Privacy in e-commerce: stated preferences vs. actual behavior, Communications of the ACM 48 (2005) 101–106.
[17] A.M. McDonald, R.W. Reeder, P.G. Kelley, L.F. Cranor, A comparative study of online privacy policies and formats, in: Proceedings of the 5th Symposium on Usable Privacy and Security, SOUPS '09, 2009, pp. 46:1–46:1.
[18] PrivacyRightsClearinghouse, The New Years Biggest Privacy Risks, 2011. <https://www.privacyrights.org/biggest-privacy-risks-2011>.
[19] FTC Staff Report: Self-Regulatory Principles For Online Behavioral Advertising, 2009. <http://www.ftc.gov/os/2009/02/P085400behavadreport.pdf>.
[20] Privacy Rights Clearinghouse. Empowering Consumers. Protecting Privacy. <http://www.privacyrights.org/>.
[21] A. Narayanan, V. Shmatikov, Myths and fallacies of "Personally Identifiable Information", Communications of the ACM 53 (2010) 24–26.
[22] V. Toubiana, H. Nissenbaum, A. Narayanan, S. Barocas, D. Boneh, Adnostic: privacy preserving targeted advertising, in: 17th Network and Distributed System Security Symposium, 2010.
[23] Interactive Advertising Bureau (IAB) and PricewaterhouseCoopers (PwC) US. Internet Advertising Revenue Report, 2012. <http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_HY_2012.pdf> (11.10.12).
[24] R. Sprague, C. Ciocchetti, Preserving identities: protecting personal identifying information through enhanced privacy policies and laws, Albany Law Journal of Science and Technology 19 (1) (2009) 91–140.
[25] J. Mayer, Web Policy, October 2011. <http://webpolicy.org/2011/10/11/tracking-the-trackers-where-everybody-knows-your-username/>.
[26] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proceedings of the National Academy of Sciences (2013).
[27] D. Butler, Data sharing threatens privacy, Nature 449 (7163) (2007) 644–645.
[28] A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets, in: Proceedings of the 2008 IEEE Symposium on Security and Privacy, SP '08, IEEE Computer Society, Washington, DC, USA, 2008, pp. 111–125.
[29] D.D. Berger, Balancing consumer privacy with behavioral targeting, Santa Clara Computer and High Technology Law Journal 27 (April) (2010) 3–61.
[30] D. Irani, S. Webb, K. Li, C. Pu, Large online social footprints – an emerging threat, in: CSE '09, International Conference on Computational Science and Engineering, 2009. vol. 3, 2009, pp. 271–276.
[31] B. Krishnamurthy, C.E. Wills, On the leakage of personally identifiable information via online social networks, in: Proceedings of the 2nd ACM Workshop on Online Social Networks, WOSN '09, ACM, New York, NY, USA, 2009, pp. 7–12.
[32] B. Krishnamurthy, C.E. Wills, Privacy leakage in mobile online social networks, in: Proceedings of the 3rd Conference on Online Social Networks, WOSN'10, USENIX Association, Berkeley, CA, USA, 2010, p. 4.
[33] B. Krishnamurthy, K. Naryshkin, C.E. Wills, Privacy leakage vs. protection measures: the growing disconnect, in: Web 2.0 Security and Privacy Workshop, 2011.
[34] C.E. Wills, M. Zeljkovic, A personalized approach to web privacy – awareness, attitudes and actions, Information Management & Computer Security 19 (1) (2011) 53–73.
[35] G.V. Lioudakis, E.A. Koutsoloukas, N.L. Dellas, N. Tselikas, S. Kapellaki, G.N. Prezerakos, D.I. Kaklamani, I.S. Venieris, A middleware architecture for privacy protection, Computer Networks 51 (16) (2007) 4679–4696.
[36] S. Bhagat, G. Cormode, B. Krishnamurthy, D. Srivastava, Privacy in dynamic social networks, in: WWW, 2010, pp. 1059–1060.
[37] C. Jackson, A. Bortz, D. Boneh, J.C. Mitchell, Protecting browser state from web privacy attacks, in: WWW '06: Proceedings of the 15th International Conference on World Wide Web, ACM, New York, NY, USA, 2006, pp. 737–744.
[38] B. Krishnamurthy, C. Wills, Privacy diffusion on the web: a longitudinal perspective, in: WWW '09, 2009, pp. 541–550.
[39] B. Krishnamurthy, C.E. Wills, Generating a privacy footprint on the Internet, in: Proceedings of the ACM SIGCOMM Internet Measurement Conference, Rio de Janeiro, Brazil, 2006.
[40] B. Krishnamurthy, C.E. Wills, Characterizing privacy in online social networks, in: Proceedings of the First Workshop on Online Social Networks, WOSP '08, ACM, New York, NY, USA, 2008, pp. 37–42.
[41] U. Shankar, C. Karlof, Doppelganger: better browser privacy without the bother, in: CCS '06: Proceedings of the 13th ACM Conference on Computer and Communications Security, ACM, New York, NY, USA, 2006, pp. 154–167.
[42] B. Krishnamurthy, D. Malandrino, C.E. Wills, Measuring privacy loss and the impact of privacy protection in web browsing, in: Symposium on Usable Privacy and Security, SOUPS '07, 2007, pp. 52–63.
[43] Privoxy Web Proxy, 2010. <http://www.privoxy.org/>.
[44] C. Canali, M. Colajanni, D. Malandrino, V. Scarano, R. Spinelli, A novel intermediary framework for dynamic edge service composition, Journal of Computer Science and Technology 27 (2012) 281–297.
[45] AdBlock Plus. <http://adblockplus.org/>.
[46] NoScript. <http://noscript.net/>.
[47] J. Samuel, B. Zhang, RequestPolicy: Increasing Web Browsing Privacy through Control of Cross-Site Requests, in: PETS '09, 2009, pp. 128–142.
[48] Ghostery. <http://www.ghostery.com/>.
[49] A. Soltani, S. Canty, Q. Mayo, L. Thomas, C. Hoofnagle, Flash cookies and privacy, in: AAAI Spring Symposium Series, 2010, pp. 158–163.
[50] C. Jensen, C. Sarkar, C. Jensen, C. Potts, Tracking website data-collection and privacy practices with the iWatch web crawler, in: Proceedings of the 3rd Symposium on Usable Privacy and Security, SOUPS '07, ACM, New York, NY, USA, 2007, pp. 29–40.
[51] M. Ayenson, D.J. Wambach, A. Soltani, N. Good, C.J. Hoofnagle, Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning, Tech. Rep., University of California, Berkeley, 2011. <http://ssrn.com/abstract=1898390>.
[52] F.T. Commission, Federal Trade Commission, Protecting America's Consumers. <http://www.ftc.gov/>.
[53] Abine, DoNotTrackMe. <https://addons.mozilla.org/en/firefox/addon/donottrackplus/>.
[54] Beeftaco. <https://addons.mozilla.org/it/firefox/addon/180650/>.
[55] Better Privacy. <http://netticat.ath.cx/BetterPrivacy/BetterPrivacy.htm>.
[56] Refcontrol. <http://www.stardrifter.org/refcontrol/>.
[57] Foundstone, Foundstone HTML5 Local Storage Explorer, 2011. <http://addons.mozilla.org/it/firefox/addon/foundstone-html5-local-storage/>.
[58] PrivacyChoice.org, Trackerblock, 2012. <http://addons.mozilla.org/en-US/firefox/addon/trackerblock/>.
[59] P. Eckersley, How Unique Is Your Web Browser? in: Proc. of the 10th international Conference on Privacy Enhancing Technologies, PETS'10, 2010, pp. 1–18.
[60] A. Cooper, RFC6462. Report from the Internet Privacy Workshop, January 2012. <http://www.rfc-editor.org/rfc/rfc6462.txt>.
[61] M. Cova, C. Kruegel, G. Vigna, Detection and analysis of drive-by-download attacks and malicious JavaScript code, in: WWW '10: Proc. of the 19th International Conference on World Wide Web, ACM, New York, NY, USA, 2010, pp. 281–290.
[62] N. Doty, E. Wilde, Geolocation privacy and application platforms, in: Proc. of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, SPRINGL '10, 2010, pp. 65–69.
[63] L. Sweeney, Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper 3, Tech. rep., Carnegie Mellon University, 2000.
[64] G. Wondracek, T. Holz, E. Kirda, C. Kruegel, A practical attack to de-anonymize social network users, in: 2010 IEEE Symposium on Security and Privacy (SP), 2010, pp. 223–238.
[65] J. Gomez, T. Pinnick, A. Soltani, KnowPrivacy. The Current State of Web Privacy, Data Collection, and Information Sharing. <http://www.knowprivacy.org/>.
[66] R. Tirtea, C. Castelluccia, D. Ikonomou, Bittersweet Cookies. Some Security and Privacy Considerations, Tech. rep., ENISA European Network and Information Security Agency, 2011. <http://www.enisa.europa.eu/>.
[67] B. Lawson, R. Sharp, Introducing HTML5, second ed., New Riders Press, 2011.
[68] K. Singh, A. Moshchuk, H.J. Wang, W. Lee, On the incoherencies in web browser access control policies, in: 2010 IEEE Symposium on Security and Privacy (SP), 2010, pp. 463–478.
[69] G. Conti, E. Sobiesk, Malicious interface design: exploiting the user, in: Proc. of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, 2010, pp. 271–280.
[70] P. Leon, B. Ur, R. Shay, Y. Wang, R. Balebako, L. Cranor, Why Johnny cannot opt out: a usability evaluation of tools to limit online behavioral advertising, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012, pp. 589–598.

[71] Revenue Science, Ponemon Study Shows Sixty – Three Percent of Consumers Always Prefer Advertising Based on Their Interests, April 2006. <http://www.audiencescience.com/press_room/press_releases/2006/20060427.asp>.
[72] A. Westin, Harris-Equifax Consumer Privacy Survey, Tech. rep., Westin, A. and Harris Louis and Associates, Conducted for Equifax Inc., 1991.
[73] A. Soltani, S. Canty, Q. Mayo, L. Thomas, C.J. Hoofnagle, Flash Cookies and Privacy, Tech. Rep., School of Information, UC Berkeley School of Law, University of California, Berkeley, 2009.
[74] C.E.Wills, M. Zeljkovic, A Personalized Approach to Web Privacy – Awareness, Attitudes and Actions, Tech. Rep., Worcester, Massachusetts, 2010.
[75] A.T.W.I. Company, Alexa Top Sites. <http://www.alexa.com/>.
[76] L. Falk, A. Prakash, K. Borders, Analyzing websites for user-visible security design flaws, in: Proc. of the 4th Symposium on Usable Privacy and Security, SOUPS '08, ACM, New York, NY, USA, 2008, pp. 117–126.
[77] S. DeDeo, Pagestats, 2006. <http://www.cs.wpi.edu/cew/pagestats/>.
[78] J. Massey, J. Frank, The Kolmogorov-Smirnov test for goodness of fit, Journal of the American Statistical Association 46 (253) (1951) 68–78.
[79] Adblocker. <https://addons.mozilla.org/sl/firefox/addon/ad-blocker/>.
[80] NAI Consumer Opt Out Protector. <http://www.networkadvertising.org/managing/protector_license.asp>.
[81] B. Krishnamurthy, I know what you will do next summer, SIGCOMM Computer Communication Review 40 (2010) 65–70.

Delfina Malandrino has received in 2000 the Laurea in Computer Science from the University of Salerno and in 2004 she received the Dottorato di Ricerca (PhD) in Computer Science at the University of Salerno (Italy). From October to December 2006 she has visited the AT&T Research Labs., New Jersey USA, working with Professor Balachander Krishnamurthy in the field of privacy protection during Web navigation. From November 2007 she is an Assistant Professor at the Dipartimento di Informatica of the University di Salerno. Her research activities mainly focus on the following research areas: distributed systems, adaptive and collaborative systems, information visualization systems, social networking, Internet traffic measurement and Benchmarking, privacy protection.

Vittorio Scarano has received in 1990 the Laurea in "Scienze dell'Informazione" (Computer Science) from the University of Salerno (Italy) and in 1995 he received the Dottorato di Ricerca (PhD) in Applied Mathematics and Computer Science at the University of Naples (Italy). In 1992, he has visited the University Eotvos Lorand in Budapest (Hungary). From 1992 to 1994, he has visited the Department of Computer Science at the University of Massachusetts at Amherst (USA) working in the team with Prof. Arnold Rosenberg in the field of parallel algorithms and architectures. He also was instructors at Mount Holyoke College (USA) in 1994. From 1995 to 2001 he was an Assistant Professor at the "Facoltá di Scienze MM. FF. e NN." (School of Mathematical, Physical and Chemical Sciences) of the University of Salerno. Since 2001 he is an Associate Professor at the Facoltá di Scienze MM. FF. e NN (School of Mathematical, Physical and Natural Sciences) of the Universitá di Salerno. He is a member of the "Dipartimento di Informatica" (Computer Science and Applications department) of the University of Salerno since 1995. His research focussed on Multimedia and Distributed Systems on the World Wide Web, covering several aspects from a theoretical perspective (P2P systems and architectures) to applications (intermediaries, cooperative systems and multimedia). Recently, his research interests are also devoted to information visualization and interactive virtual environments.