Fine-grained control over tracking to support the ad-based Web economy

Jagdish Prasad Achara, INRIA Grenoble
Javier Parra-Arnau, INRIA Grenoble
Claude Castelluccia, INRIA Grenoble

The intrusiveness of Web tracking and the increasing invasiveness of digital advertising have raised serious concerns regarding user privacy and Web usability, leading a substantial chunk of the populace to adopt ad-blocking technologies over the last years. The problem with these technologies, however, is that they are extremely limited and radical in their approach, and completely disregard the underlying economic model of the Web, in which users get content free in return for allowing advertisers to show them ads. Nowadays, with around 200 million people regularly using such tools, said economic model is in danger.

In this paper, we investigate an Internet technology that targets users who are not in general against advertising, accept the trade-off that comes with the "free" content, but —for privacy concerns— they wish to exert fine-grained control over tracking. Our working assumption is that some categories of web pages (e.g., related to health or religion) are more privacy-sensitive to users than others (e.g., about education or science). Capitalizing on this, we propose a technology that allows users to specify the categories of web pages that are privacy-sensitive to them and block the trackers present on such web pages only. As tracking is prevented by blocking network connections of third-party domains, we avoid not only tracking but also third-party ads. Since users continue receiving ads on those web pages which belong to non-sensitive categories, our approach may provide a better point of operation within the trade-off between user privacy and the Web economy. To test the appropriateness and feasibility of our solution, we implemented it as a Web-browser plug-in, which is currently available for Google Chrome and Mozilla Firefox. Experimental results from the collected data of 746 users during one year show that only 16.25% of ads are blocked by our tool, which seems to indicate that the economic impact of the ad-blocking exerted by privacy-sensitive users could be significantly reduced.

Additional Key Words and Phrases: user privacy; Internet economy; ad-blocking

1. INTRODUCTION

According to a recent PageFair report [15], 45 million Americans (16% of online users) and 77 million Europeans had installed ad-blockers[1] as of the second quarter of 2015. All this globally accounts for 21.8 billion dollars worth of blocked ads over that year. Since ads fuel the free content and services over the Web, this monetary loss puts the ad-based economic model of the Web under threat. The nature of this threat has become more devastating over time since, in the early days of online advertising, the ad industry ignored this potentially dangerous situation in favor of immediate benefits.

Several studies have investigated the reasons why Web users block ads [15; 39; 9; 14; 11; 17; 21; 34; 45]. According to those surveys, some simply do not want to receive advertising, others block ads because they are annoying and degrade their browsing experience, while another chunk of the populace finds the underlying tracking invasive and perceive advertising as a source of privacy and/or security concerns. The reasons behind ad-blocking, however, are not mutually exclusive and those reports also find users who decided to start blocking ads due to a combination of the reasons above.

The radical design choices offered by ad-blockers is another reason why ad-blocking has become a threat to the Web economy today. A scrutiny of current ad-blockers further reveals that they do not take into account (1) the economic impact of ad-blocking, and (2) the social and economic benefits of non-intrusive and rational advertising. This

[1] All tools that eventually end up blocking ads even though they are initially designed for some other purpose, such as privacy and transparency.

demands better tools that cater to users' concerns, while at the same time are designed to ensure that their economic impact does not get sidelined.

In this work, we address the evoked concerns by proposing a three-dimensional approach that combines user privacy, ad-blocking and its economic impact on the Internet. We target users who are not against ads and accept the trade-off that comes with the "free" content. However, for privacy concerns, they may wish to exert a fine-grained control over tracking and hence block some ads.

Our working assumption is that users may consider their visits to some web pages (for example, religion- and health-related pages) more privacy-sensitive than to others (e.g., news and sports) [36; 24]. According to this assumption, our approach allows users to specify the categories of web pages where they do not want to be tracked on, and thus do not want to receive third-party ads[2]. Unlike current ad-blockers, our technology enables users to continue receiving ads on categories of web pages that they consider as non-sensitive.

To test the viability of our approach, we have developed *MyTrackingChoices*[3], a Web-browser plug-in for Google Chrome and Mozilla Firefox. When users of our technology browse the Web, the tool categorizes the visited web pages on the fly and, depending on the users' choices, the network connections of the third-party domains present on those pages may be blocked. As tracking is prevented by blocking third-party network requests, the proposed technology may avoid not only tracking but also third-party ads. We would like to stress that, in contrast to current ad-blockers, our tool does not preclude those ads which are served directly by a publisher.

Contributions. Next, we summarize the major contributions of this work:

— We review the current ad-blocking technologies in terms of both their design choices and the ensuing impact on the Web economy. Our findings reveal that those technologies are unappropriate to users and the Internet economic model. We also show that the current self-regulatory initiatives by the ad industry and various other organizations is insufficient, as users' choices are not guaranteed to be enforced.

— Building on the lessons learned, we propose a tool that allows users to exert fine-grained control over tracking and advertising. Being user-centric, the proposed system ensures that privacy and ad preferences are enforced on the user side. A fine-grained control, on the other hand, enables users to find a better trade-off between their privacy and the economic model that currently sustains the Web.
Our solution permits users to choose those categories of web pages which are privacy-sensitive to them. MyTrackingChoices technically enforces their choices by blocking the network connections of third-party domains (and thus, the ads delivered through third-party domains) which are present on those pages. Our approach is very much in line with a recent study [40] on users' preferences for Web tracking, which concludes that the majority of the surveyed users (1) "commonly base their tracking preferences on specific properties of first-party websites such as the topic of the site"[4].

— To test the appropriateness and feasibility of our approach, we propose a system architecture and implement it as a Google Chrome extension and a Mozilla Firefox plug-in. We conduct an extensive evaluation of our proposal in terms of categorization efficiency, usability and performance. We find that our categorization algorithm

[2] Such ads are delivered by domains other than the one browsed by a user.

[3] https://chrome.google.com/webstore/detail/mytrackingchoices/fmonkjimgifgcgeocdhhgbfoncmjclka, https://addons.mozilla.org/en/firefox/addon/mytrackingchoices/

[4] The cited work describes the first investigation of users' tracking preferences carried out in the context of their own browsing histories. The authors collected browsing histories from and interviewed 35 people about the perceived benefits and risks of online tracking. Among other questions, the respondents were asked to review a number of aspects related to most popular ad blockers and anti-tracking tools.

performs reasonably well in all topic categories and especially in those which were considered sensitive by the participants of our experiments.

— Last but not least, we report experimental results based on the pseudo-anonymous data collected from 746 users of our tool during the period April 2016 - April 2017. Remarkably enough, our experimental analysis shows that the percentage of blocked ads is just 16.25, which suggests that the economic impact of the ad-blocking exerted by privacy-sensitive users could be largely diminished.

2. EXISTING AD-BLOCKERS: PROBLEMS AND PERSPECTIVES

Due to the proliferation of intrusive[5] and privacy-invasive ads, ad-blockers have become extremely popular over the last years. These Web technologies can be grouped into two classes: *ad-blockers* and *anti-trackers*. Ad-blockers are technologies which merely block ads, while anti-trackers, although eventually block ads too, their aim is to provide higher-level functionalities such as privacy protection and transparency. Some examples of ad-blockers include AdBlock [32] and AdBlock Plus [1], whereas Ghostery [3], Disconnect [2], PrivacyBadger [5] are examples of anti-trackers. In terms of functionalities, ad-blockers in the first class block all ads whereas anti-trackers allow users to block a particular tracker or category of trackers (e.g., related to analytics, privacy, advertising) or trackers on a per domain basis. Considering this distinction in terms of their functionalities and objectives, in the remainder of this section we make the distinction between them.

Third-party ad-delivery and ad blockers operation. There exist two ad-delivery models. In *first-party advertising*, advertisers negotiate with publishers the ads that will be shown to visitors of their websites; the ads served this way are delivered by publishers and are essentially untargeted. In *third-party advertising*, on the other hand, advertisers engage the services of an ad platform, which is responsible for displaying their ads on the publishers' sites.

In this latter model, the ad-delivery process begins with publishers embedding in their sites a link to the ad platform they want to work with. The upshot is as follows: when a user retrieves one of those websites and loads it, their browser is immediately directed to all the embedded links. From these links, the ad platform may track this user's visit and display the ads of the advertisers it partners with.

Within the third-party ad model, real-time bidding (RTB) is the dominant technology with 74% of programmatically purchased advertising [46]. The main difference with respect to traditional ad platforms is that the decision on which ad should be displayed to a user in a given ad space no longer depends on the ad platform. Rather, advertisers are allowed to bid for each user and ad space depending on their specific requirements.

In the ad-blocking game, anti-trackers and ad blockers act as firewalls between a user's browser on the one hand, and the ad platforms and tracking companies on the other. Specifically, ad blockers operate by preventing those HTTP requests which are made when the browser loads a web page, and which are not originated by its publisher. These requests are commonly referred to as third-party network requests, and blocking them implies blocking all third-party ads.

Drawbacks of current ad-blockers and anti-trackers. The current ad-blocking technologies, however, constitute a radical approach in its bid to counter invasive advertising: users can only decide either to block or allow all ads. Besides, the fact such technologies are configured to block all ads by default (and do not let users specify their own choices) might suggest that their developing companies did not consider their economic impact on the Web.

[5] Ads are considered to be intrusive if they hide content or if they pop up randomly on the screen making user browsing experience frustrating.

Anti-trackers give users the option to decide by which trackers or category of trackers they do not want to be tracked. However, we argue that most users are not concerned with the trackers but with another dimension of tracking, namely, on which web pages they do not want to be tracked. In this respect, anti-trackers let users decide to block trackers on a per domain basis, i.e., users can whitelist or blacklist a specific domain. However, we believe that such domain-level granularity is not an appropriate approach for three main reasons:

(1) Given the huge number of domains, it is almost impossible for users to determine and pre-define all the domains where they do not want to be tracked.
(2) Some domains can host web pages belonging to different categories, for example, belonging to both sensitive (health, religion, etc.) and non-sensitive (sports, science, etc.) categories. Those domains belonging to the "news" category, e.g., `cnn.com` or `bbc.com`, are good examples since they usually include web pages from a variety of categories (sports, economy, politics, health, travel, religion, etc.). Therefore, except for some domains which have all web pages belonging to a same category, it makes more sense to block trackers based on the category of the web page and not on a per domain basis. In fact, page-granular (and not domain-granular) blocking makes more sense both from users' privacy point of view, as recently shown by [40], and with respect to reducing the economic damage experienced by publishers. This is partly justified by the fact that, for a given domain, ads would eventually be blocked only on web pages belonging to sensitive categories chosen by users and not on the whole domain.
(3) Blocking trackers based on the category of web pages (instead of configuring it for each and every domain) makes it easier for users to configure as they just need to select once the categories of web pages that are sensitive to them. As the number of web page categories is necessarily limited, these pre-defined categories can be made available to the user when the tool is installed and user choices can be respected from the very beginning.

With the exception of PrivacyBadger [5], another issue with the existing ad-blockers and anti-trackers is that, in order to block trackers or ads, they rely on black-lists manually maintained by their developers or, in some cases, by user communities. The use of these lists by AdBlock Plus [1], currently the most popular among ad-blocking tools, stirred controversy for accepting money from some ad companies to whitelist them [25; 6]. Furthermore, it is very cumbersome to maintain these lists as new trackers and/or ad companies keep appearing in the market and already existing companies keep introducing new mechanisms to deliver ads.

Self-regulatory initiatives by the ad industry are not appropriate. Several self-regulatory initiatives [13; 10] from the advertising industry have been proposed to address some of the concerns behind ad-blocking such as Web usability and performance. These efforts envisage to improve the ad experience for users but none of them returns control to users. Moreover, these initiatives from the ad industry do not take into account the privacy concerns arisen from ad-blocking, even though various studies confirm that a non-negligible number of users block tracking and ads just to protect their online privacy [15; 14; 11]. The LEAN program from the Interactive Advertising Bureau (IAB) [4], one of the biggest advertising organizations with over 5 500 ad companies and publishers, discusses several aspects of non-invasive advertising in a guide to actively fight ad-blocking, but it is not in terms of privacy [13]. In this same line, the "acceptable ads manifesto" includes five points to improve the ad-experience, in an attempt to make ads acceptable by users. However, none of these five points tackles the privacy risks [10] very often caused by persistent tracking and invasive advertising.

Table I: Top-level interest categories.

adult	economics	hobbies & interests	politics
agriculture	education	home	real estate
animals	family & parenting	law	religion
architecture	fashion	military	science
arts & entertainment	folklore	news	society
automotive	food & drink	personal finance	sports
business	health & fitness	pets	technology & computing
careers	history	philosophy	travel

Other examples of self-regulatory initiatives are "Your Online Choices" (from a group of European organizations) [7] and DNT (from the World Wide Web Consortium) [16]. The former lets users block ads tailored to their web browsing interests but users can never be sure if their choices are actually honored and if they are still being tracked. These initiatives are insufficient as users' choices are not technically enforced at their side. Similarly, DNT allows users to notify websites and the ad industry if they want to stop being tracked through third-party cookies. But then again, the problem of enforcing users' preferences over tracking and advertising still persists.

3. MYTRACKINGCHOICES

As we have discussed in Sec. 2, the existing ad-blocking technologies and self-regulatory initiatives are not appropriate to deal with the current threat to the ad-based economic model of the Web. Nowadays, there is an urgent need for tools that give users more fine-grained control over tracking and thus, third-party ads, as recently identified by the first in-depth investigation of users tracking preferences [40].

3.1. New approach to anti-tracking

We propose a different approach as a solution to the afore-described concerns. Our approach targets users who are not against ads but wish to block them due to privacy reasons. The proposed system design relies on the assumption that most people do not want to be tracked on "sensitive" websites (for example, related to religion and health), but would accept to be tracked and receive ads on less sensitive ones (such as news, sport) in order to support the content provider.

The ultimate aim of this technology is to sustain the ad-based economic model of the Web and, hence, it is by design in contrast with the rationale behind existing ad-blockers. The idea is to let users choose the categories of web pages that are privacy-sensitive to them (for example, health and religion) and exclusively block the respective trackers and/or ads on those web pages. The granularity of selecting where users accept ads gives it an advantage over other ad-blockers in various respects as mentioned in Sec. 2. Thanks to our approach, users can continue receiving profile-based targeted ads[6] on categories of web pages they accept to be tracked.

Our approach does not follow an "all-or-nothing" policy, in contrast to current ad-blockers. We allow users to select the categories of web pages that they consider sensitive and do not want to be tracked and receive ads on. Furthermore, users can be even more selective in the sense that they can block a web page without blocking the entire category. For instance, if a particular web page hosts intrusive ads that belongs to a non-sensitive category, users can selectively penalize it while allowing ads on other web pages in the same category. This feature may encourage publishers not to include intrusive ads on their web pages.

[6] As users are okay to be tracked on some categories of web pages. They still can receive useful ads targeted to those interest-categories of users.

It is worth noting that our proposal is not a replacement of the current initiatives proposed by the online advertising industry, but a complement to them. While these initiatives aim at improving the quality of the ads delivered to users, we permit users to decide where they want to accept tracking and third-party ads. We emphasize that users who outrightly reject all ads are not considered. Other economic models (e.g., subscription-based access to Web content) must be put in place for such users [18].

MyTrackingChoices is a tool that implements this approach and is described in more details in the following section.

3.2. Implementation details

MyTrackingChoices has been implemented as a Google Chrome extension[7] and a Mozilla Firefox plug-in[8], and are both available for download. In terms of functionalities, our tool permits users to block trackers (and thus, third-party ads) based on pre-defined categories of web pages or on a per web page basis. If users do not agree with the categorization of a web page by MyTrackingChoices, they may change the category of that web page. In an attempt to bring transparency over Web tracking, users can also learn the third-party domains present on a web page.

MyTrackingChoices currently supports web pages in English, French, Spanish and Italian. This is due to the limitation of our current *categorizer* module (described next) but we plan to extend support for more languages in the near future.

The system architecture of the proposed technology consists of three main modules, each of them performing a specific task. These modules are the *categorizer*, the *policy module*, and the *blocking module*. Next, we specify the conceptual design and fundamental operational structure of a practical implementation of this technology. In the coming sections, the terms "web page" and "URL" will be used interchangeably.

3.2.1. Categorizer. This module classifies the pages visited by users into a predefined set of topic of interests. The module employs a 2-level hierarchical taxonomy, composed of 32 *top-level categories* and 330 *bottom-level categories or subcategories*. For the sake of usability, in the current version of our technology we display and let users interact only with top-level categories. However, depending on the feedback of users, e.g., if they find the top-level categorization too coarse, we can further allow users to interact with *bottom-level categories*. A list with the top-level categories can be found in Table I.

The categorization algorithm integrated into our system is partly inspired by the methodology presented in [33] for classifying non-textual ads into interest categories. The algorithm also builds on the taxonomy available with the Firefox Interest Dashboard plug-in [12] developed by Mozilla.

Our categorizer relies on two sources of previously-classified data. First, a list of URLs, or more specifically, domains and hostnames, which is consulted to determine the page's category. Here, it is worth noting that this list of URLs only contain domains that have all web pages belonging to a particular category. For instance, in the current version of our plug-in the domain `techcrunch.com` is mapped to "technology & computing" and "news". Second, a list of unigrams and bigrams [38] that is used when the URL lookup fails. The former list is justified by the fact that a relatively small part of the whole Web accounts for the majority of the visits. [29]. It is evident that precategorized lookup requires few computational resources on the user's browser and can be more precise. The latter list, on the other hand, is justified as a fall-back and allows us to apply common natural-language heuristics to the words available in the URL, title, keywords and content of the web page.

[7] https://chrome.google.com/webstore/detail/mytrackingchoices/fmonkjimgifgcgeocdhhgbfoncmjclka

[8] https://addons.mozilla.org/en-US/firefox/addon/mytrackingchoices

To build the first list of URLs, we incorporate Alexa.com's 500 top Web sites for almost each of the top-level categories. The list also includes the domains and hostnames (around seven thousand) classified by Mozilla's Interest Dashboard plug-in [12].

On the other hand, in the list of unigrams and bigrams, we have approximately 76 000 entries for English language. Three additional lists of unigrams and bigrams, although with a fewer number of entries, are also available for French, Spanish and Italian. To compile all these words lists, we have built on the following data:

— a refined version of the categorization data provided by the Firefox Interest Dashboard extension;
— a subset of the English terms available at WordNet 2.0 [41] for which the WordNet Domain Hierarchy [37; 22] provides a domain label;
— a subset of the terms available at the WordNet 3.0 Multilingual Central Repository [31], to allow the categorization of sites written in the aforementioned languages;
— and the synset-mapping data between the versions 2.0 and 3.0 of WordNet [27].

The categorizer module resorts to word lists only when the hostname and domain are not found in the URL list. When this happens, the algorithm endeavors to classify the page by using the unigrams and bigrams extracted from the following data fields: URL, title, keywords and content of the page. Depending on the data field in question, the categorizer assigns different weights to the corresponding unigrams and bigrams. In doing so, we can reflect the fact that those terms appearing in the URL, the title, and especially the keywords specified by the publisher (if available), are usually more descriptive and explanatory than those included in the body of the page.

As frequently done in information retrieval and text mining, our web page classifier also relies on the Term Frequency-Inverse Document Frequency (TF-IDF) model [44]. Said otherwise, we weigh the resulting category(ies) based on the frequency of occurrence of the corresponding unigrams and bigrams in the web page, and on a measure of their frequency within the whole Web[9]

Based on the categorization using words, we have a score for each category depending on the algorithm above described. To select the categories that best represent a web page, we select all the categories that have a score larger than a threshold value, where threshold (T) is defined as

$$T = \alpha \times (\text{maximum score} - \text{average score}),$$

where α is a constant and its value is chosen based on how much weight we want to give to accuracy versus completeness (false positives versus true negatives). In our current implementation, we choose α equal to 0.3. If there are less than three such categories, the web page is classified into all these categories. However, if there are more than three such categories, we pick only three top categories and categorize the web page into them. Thus, the maximum number of categories a page may belong to is upper bounded by 3.

For the sake of computational efficiency, the algorithm caches the categories derived from the user's last 500 visited pages. This way, when the user re-visits one of those pages, the interest-categories of these pages are obtained directly without needing to go through the whole process of categorization described above.

[9] We estimate the frequency of unigrams and bigrams with the help of the Corpus of Global Web-based English [26], which contains words from blogs and other Web-based materials such as newspapers, magazines and company websites.

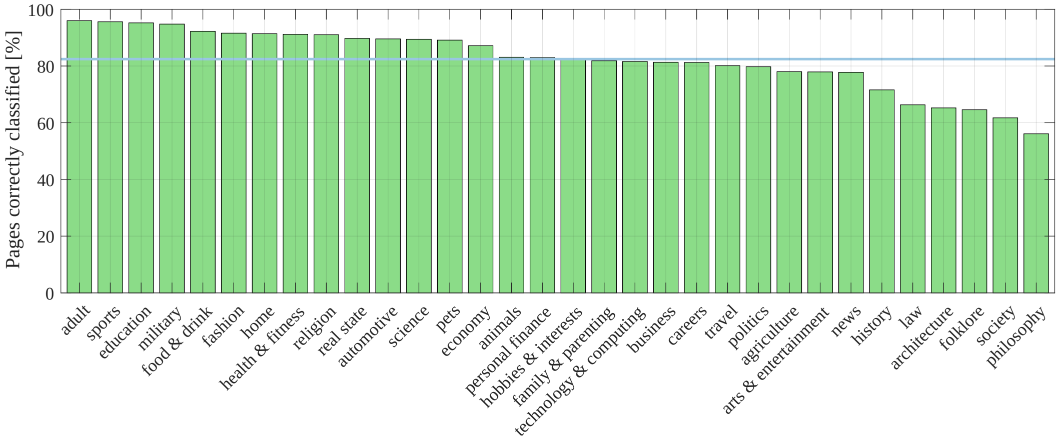

Fig. 1: Percentage of pages correctly classified per category in our evaluation of the categorization algorithm.

In terms of storage, the whole list of unigrams, bigrams and their corresponding IDF values occupies approximately 1 megabyte in compressed format. We believe this is an acceptable overhead to the plug-in download size.

To evaluate the performance of our categorizer, we relied on the data set available at the Open Directory Project (ODP) [10], which provides a list of web sites together with its classification into a given taxonomy. The data set we employed in our empirical evaluation is a subset of these web sites.

The methodology that we followed to assess our categorization algorithm is summarized next.

— We first found a correspondence between the top-level categories of our taxonomy and those of ODP. In those cases where it was not possible to find a mapping at the top-level, we resorted to the second and third levels of the taxonomy of ODP.
— For each of our 32 categories, we collected one thousand web pages classified accordingly by ODP.
— For each of the 32 000 pages collected, we executed our classification algorithm and checked whether the category with the highest score coincided with the topic available in the ODP data set. To this end, we disabled the URL lookup function, and the algorithm thus based its decision only on the URL, title, keywords and content of the web page. We proceeded this way because we aimed at testing the non-evident part of our algorithm, in which URLs are not categorized manually.

The results are shown in Fig. 1 and the main conclusion that can be drawn is that our categorization algorithm performs quite satisfactorily. More specifically, the percentage of pages correctly classified by the categorizer was 82.42% in average overall, being 95.98%, 91.15% and 91.02% for “adult”, “health & fitness” and “religion”, respectively. As we shall see in Sec. 4.1, these three categories will be the most sensitive categories to the participants of our experimental evaluation.

Lastly, we would like to point out that users may also classify manually the page being visited when they disagree with the results provided by our algorithm. In our experiments, users reported 31 pages with inaccurate classification during one year approximately.

[10] http://www.dmoz.org

3.2.2. Policy module. This module is responsible for applying the blocking policies defined by users. Policies can be defined either on the content category or on a per page basis.

To simplify the implementation, our extension allows users to specify only per category-blocking policies. That is, instead of defining categories of web pages where tracking should be allowed and categories of web pages where tracking should be blocked, we just enable the latter blocking declaration.

We would like to emphasize that the particular choice of an anti-tracking policy, and the corresponding privacy benefit, are absolutely dependent on users' own perception regarding web-content sensitivity. When a user installs an ad blocker or anti-tracker, all third-party tracking is prevented and hence no profiles are built from their browsed pages. In our case, if a page category is declared as sensitive, our tool blocks all tracking on it, either. Consequently, blocking those categories regarded as sensitive by a user means this user feels safe while browsing and being tracked on the non-sensitive categories. Otherwise the user would declare such particular non-sensitive category/ies as sensitive so as to avoid the tracking on them.

A configuration panel is shown to users that allows them to specify their anti-tracking preferences per category. The panel is presented after having installed the extension and before the extension starts functioning and, by default, no categories are blocked. The decision of which categories should be blocked can also be made at a later stage by clicking on the "configure your tracking choices" option in the popup page.

In contrast to the per-category policies, users are enabled to configure both block and allow per URL policies. This is to allow users to be more granular if they are occasionally not satisfied with per category-blocking policies. For example, in a scenario where a user has blocked a category of web pages but wants to support a specific web page in that category by receiving ads.

The operation of this module is described next. When a user visits a page, the module waits for the categorizer to send the topic category of the web page browsed. Being equipped with the URL as well as the category, the module decides whether the network connections of third-party domains should be blocked or not. We emphasize here that in case of a conflict between per web page and per category based policy, per URL based policy prevails. In the event that a web page is categorized in more than one category, if any of the category is selected by the user to be blocked, the third-party network connections on that web page are blocked.

3.2.3. Blocking module. This module is responsible for blocking network connections of third-party domains if a user has selected to block them. There are two main tasks. First, finding the third-party domains present on a page, and then blocking them. To find third-party domains, we just need to check if the domains of network connections match with the domain the user typed in the address bar. If it is not the same, then such domains can be considered as third-party domains.

The next task is to block the network connections to such domains. This task is not trivial, since blocking network connections of all third-party domains may break the functionality of some web pages. The reason is due to the fact that some web pages download useful content from other domains (a different domain belonging to the first-party domain or a content provider or some other domain). To address this problem, existing ad-blockers keep the list of domains that deliver ads or are tracking users. More specifically, this list of domains is also accompanied with some regular expressions that are used to pinpoint only some network connections from those domains.

MyTrackingChoices follows a different approach to ascertain which third-party network connections should be blocked so that the functionality of the page being visited is not broken. Instead of gathering and storing all tracking and advertising domains —this process may be cumbersome and highly dynamic—, we maintain a list of domains that are essential for the functionality of a web page. This list includes popular service providers such as Google and Facebook, as well as the most common content-delivery networks[11]. In our experience, such a list is smaller and easier to maintain than the list of tracking and advertising domains. Therefore, we only store the list of such domains [20], and block all other domains which are not included in said list. These domains are identified as "trackers" based on a heuristic described below.

We classify a third-party domain as a "tracker" if it is present on three or more different domains that a user visited in the past. This implies that the extension becomes fully functional only after a user visits a couple of web pages after installation. Our heuristic is employed so that we do not need to maintain a list of third-party domains that are specific to a first-party domain to deliver content. For example, `lemonde.fr` uses `lemde.fr` domain to deliver content to their web pages. However, this domain is only specific to `lemonde.fr` and is probably used by Le Monde only for this purpose. Consequently, such domains are not classified as "trackers". As we do not need to keep the list of such domain specific third-party domains, this helps us to keep our list of useful domains reasonably small.

Finally, although our plug-in is configured to classify trackers on the basis of three observations on distinct domain sites, we would like to stress that users themselves may ultimately invalidate this classification by reporting a page malfunctioning at any time. This heuristic provides some flexibility to the developed tool while it permits maintaining a tractable list of allowed third-party domains as non-tracking entities.

4. ECONOMIC IMPACT

This sections aims to evaluate to what extent our approach could contribute to decreasing the impact of ad-blocking on the Web economy. The evaluation is conducted on the basis of the data collected from users of MyTrackingChoices. In the sequel, we first describe our data set and the ethical considerations regarding its collection and use. Afterwards, we analyze the impact of our technology in terms of tracking ubiquitousness, user anti-tracking policies, blocked pages and blocked ads per category, and attempt to estimate the actual economic cost of our users on the Web and how the Internet would benefit from this technology if it was massively adopted.

Data set. Our analysis has been carried from the browsing data and ads of 746 users of the Web-browser extension[12]. It is worth noting that these data correspond to users who downloaded *MyTrackingChoices* directly from the Chrome Web Store. Said otherwise, they were not specifically recruited to participate in our experiment.

Our series of experiments were run from April 2016 to April 2017, and allowed us to capture 1 624 142 page visits and 110 642 ads. It is important to mention that searches on Google, Yahoo and any other Web-search engine, as well as page refreshes in Gmail and other e-mail services, are each counted as one page visit. However, because each URL is hashed (instead of the domain name) and the URL of a search or an e-mail page contains the searched terms and some sort of user identifier, it is impossible for us to know the particular service or website visited. The upshot is that among the aforementioned number of page visits, 479 411 of them were unique hashed URLs. This large number of unique page visits might suggest that a portion of them correspond to Web searches and e-mail services, which would be, on the other hand, consistent with

[11] This list is updated on the basis of the broken pages reported by users.

[12] We consider one installation as one user. However, it is technically possible for a user to uninstall the extension and install it again. We assume that users did not do this.

the fact that, among Alexa top 500 sites[13], search engines and e-mail pages lead the ranking of most popular websites.

The data set used in our analysis, however, was preprocessed so as to retain users with a sufficient browsing activity. In particular, through a process of outlier removal we decided to get rid of those users who visited less than 20 pages and never configured their anti-tracking preferences, either on a per category or on a per web page basis. Those users who did not set any blocking preference at all have not been considered in this study because our tool serves no purpose in this case. The choice of a minimum browsing of 20 web pages is somewhat arbitrary but the intuition behind it is not to take into account the preferences of users who did not considerably use the extension. Such users might further distort the ensuing analysis.

In terms of usage, our set of users used MyTrackingChoices from 1 day to a maximum of 369 days. Even though it is possible for users to continue using the extension and just opt-out of data upload, we consider that they stopped using our tool when no data was received from them. Among our 746 users, there were 4.29 % of them who used our extension for less 1 day or less, on average being 37.69 days.

Ethical considerations regarding data collection. The collected data is considered to be pseudo-anonymized because it does not contain any information that can be directly used to re-identify users. It contains hashes and categories of visited pages along with third-party domains and URLs of iframes present on those pages. As a part of responsible disclosure, users were informed that the tool is developed by Inria as a research project and that the collected data would be utilized to study the impact of MyTrackingChoices on the Web. We allow users to opt-out of data collection as well as to delete previously uploaded data to our servers. Furthermore, we assured users not to make an attempt to re-identify them and also not to share the data with any outside party. The privacy policy[14] of our tool clearly specifies all these details regarding data collection, storage, processing and sharing.

4.1. Analysis

To ensure a proper functioning of the extension, we ensured that users uninstall other ad-blockers and/or disable the blocking functionality provided by other anti-trackers.

Since we aim to investigate the impact of our technology on the Web economy, our analysis focuses on statistics aggregated over all users of MyTrackingChoices. As a result, most of the results shown next examine this impact at the level of Web-page category or publisher type.

We begin our analysis by looking at the number of users who exercised fine-granular control. We find that 67.83% of users accepted the presence of third-party domains and hence tracking and hosting ads in some categories of web pages that they do not consider much sensitive. On average, a user blocked 13.75 categories out of a total of 32 categories. The median of the number of blocked categories is 11. These bare statistics appear to suggest that the impact on the Web economy due to ad-blocking might be reduced if users are provided with such fine-grained choices. However, we need to keep in mind that such a solution is intended for users who block ads because of privacy reasons, who in principle are not against ads and do not want to be tracked on some categories of web pages that they regard as sensitive.

We also notice that 32.17% of users regarded all categories as sensitive and consequently decided to block all network connections of third-party domains. These are probably the users that are either against tracking at all or the ones who wish to prevent any form of advertising; as a matter of, blocking trackers on all visited pages

[13] http://www.alexa.com/topsites

[14] Available at https://myrealonlinechoices.inrialpes.fr/privacy_policy.html

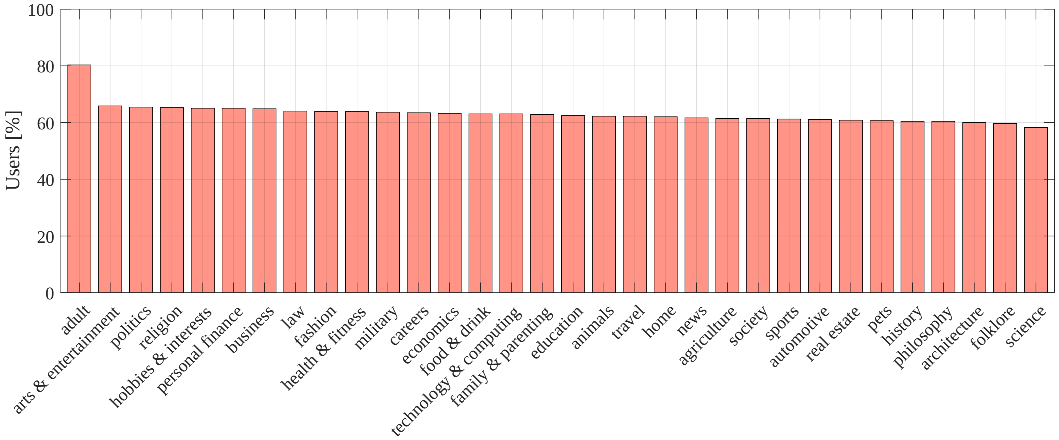

Fig. 2: Percentage of users blocking a category.

implies blocking a significant portion of ads, since a common way to deliver advertising today on the Web is through third-party domains present on web pages. Because such users (i.e., those who wish to block all page categories) are not employed a fine-granular control over tracking, the adoption of our tool in such cases will not contribute to support the ad-based Web economy.

Next, we explore the categories of web pages that are most or least blocked by our 746 users. This will enable us to know the categories that are the most sensitive to users —where most users do not want to be tracked on— and the categories that users care the least with respect to tracking. With this information, our tool could try to include the categories of web pages to block by default and let user tweak them if they are not satisfied.

Fig. 2 represents the blocked categories in the order of the most blocked to the least blocked. From this figure, we observe that the four categories most affected by blocking are “adult”, “arts & entertainment”, “politics” and “religion”, which precisely, and with the exception of “arts & entertainment”, are content categories regarded as *sensitive* by the European Data Protection Law [8]. The three least blocked topics, on the other hand, are “architecture”, “folklore” and “science”. Based on this straightforward observation, it follows that users of MyTrackingChoices are mostly concerned about being tracked on pages with sensitive content [8]. Consequently, it makes sense to provide them with such a category-based control over tracking.

Having analyzed the categories of pages that were most and least affected by the adoption of our tool, now we proceed to explore users’ browsing habits, and in particular the category of web pages they browse the most/least. This will provide insight on the impact of the proposed technology on the Web.

Fig. 3 portrays the per-category distribution of all pages browsed by the users of our data set, as well as the percentage of blocked versus allowed pages in each category. It is important to emphasize that, in our computation of said distribution and all our results of this subsection, we considered the classification errors of our categorizer, estimated for a sample of ODP in Sec. 3.2.1. Hence, the results shown reflect the inherent errors of our categorization algorithm in the estimation of the pages and ads that were actually blocked by our plug-in. In order to detect any bias in such data, on the other hand, we checked it with other studies and the observed pattern seems to resemble well [30; 35].

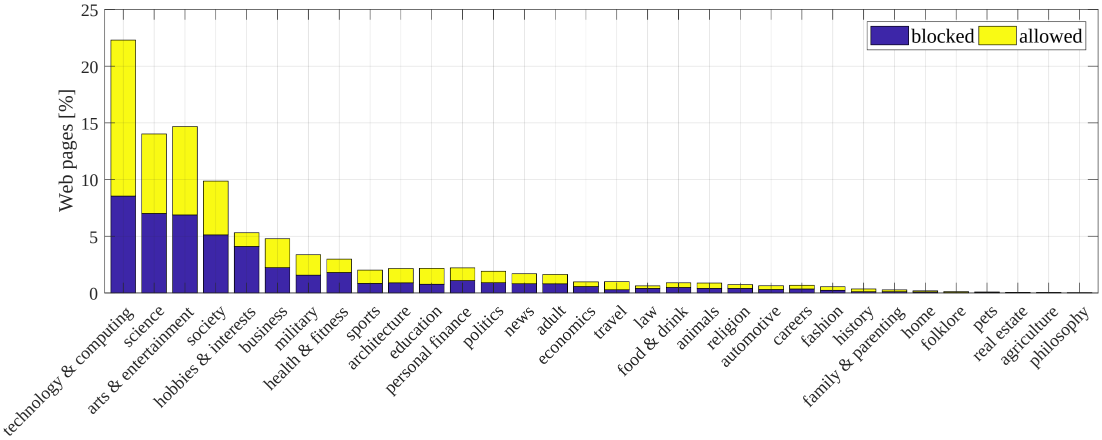

Fig. 3: Distribution of blocked and allowed visited pages per content category.

We found that the top-5 most browsed categories were “technology & computing”, “science”, “arts & entertainment”, “society” and “hobbies & interests”. Here it is worth noting that the pages belonging to these 5 categories accounted for 66.17% of the total browsing activity. However, none of those categories —with the exception of “arts & entertainment”— were included in the top-5 most blocked categories, as shown in Fig. 2. This partly explains why there were more allowed than blocked pages in the top-5 categories of Fig. 3.

We observe that there were more blocked than allowed web pages in Fig. 3 than for the categories that are the most blocked in Fig. 2. Since the most blocked categories are least browsed and the most browsed categories are the least blocked, the resulting overall impact on the Web economy is clearly less damaging. Having said this, one of the most relevant results of our analysis is that 48.13% of the browsed pages were actually blocked, which represents a significant reduction in the impact of tracking blocking —and hence third-party ad blocking— when compared to the strategy proposed by current ad blockers and anti-trackers, which merely aims to prevent tracking across all browsed pages and eliminate network-based advertising.

Until now, we have examined the impact that fine-grained control over tracking might have on the Web. The obtained results appear to suggest that MyTrackingChoices might diminish the effect of anti-tracking on the prevalent economic model of the Internet. In the sequel, we shall explore the actual effect of such anti-tracking on advertising. We shall proceed by examining the source URLs of all iframes included in the web pages browsed by users, and by computing how many of those iframe-based ads were blocked by our tool. Our aim is to estimate how the anti-tracking policies specified by our set of users impacted advertising and hence the economy of the Web.

Our estimation leverages the fact that iframes are generally used to deliver *display ads* from third-party domains directly into a web page. It is worth stressing, however, that our plug-in cannot detect non-iframe based ads which may include textual and video ads, as well as those ads delivered directly by the publisher.

In our data set, we found 965 distinct domains delivering iframes to our users. However, since iframes do not necessarily contain ads, we first filtered out those iframes which were employed to deliver ads. This filtering was done on the basis of the list of advertising domains from the Mozilla Focus project and their partner Disconnect [19]. After this filtering, the number of distinct ad domains and the number of iframes (ads served by such domains) became 204 and 110 642, respectively. Table II shows a list

Table II: Top 30 ad domains that delivered ads through iframes

doubleclick.net 21.64%	2mdn.net 2.20 %	doubleverify.com 0.97 %
googlesyndication.com 17.54%	openx.net 2.14%	criteo.net 0.95%
criteo.com 5.53%	augur.io 2.13%	betrad.com 0.93%
outbrain.com 3.89%	exoclick.com 1.87%	intermarkets.net 0.85%
rubiconproject.com 3.67%	amazon-adsystem.com 1.56%	serving-sys.com 0.80%
adnxs.com 3.19%	weborama.fr 1.46%	adsrvr.org 0.77%
imrworldwide.com 3.10%	krxd.net 1.41%	casalemedia.com 0.53%
pubmatic.com 3.05%	atwola.com 1.32%	flashtalking.com 0.47%
bluekai.com 2.66%	demdex.net 1.22%	mathtag.com 0.46%
gemius.pl 2.26%	adverticum.net 1.16%	advertising.com 0.45%

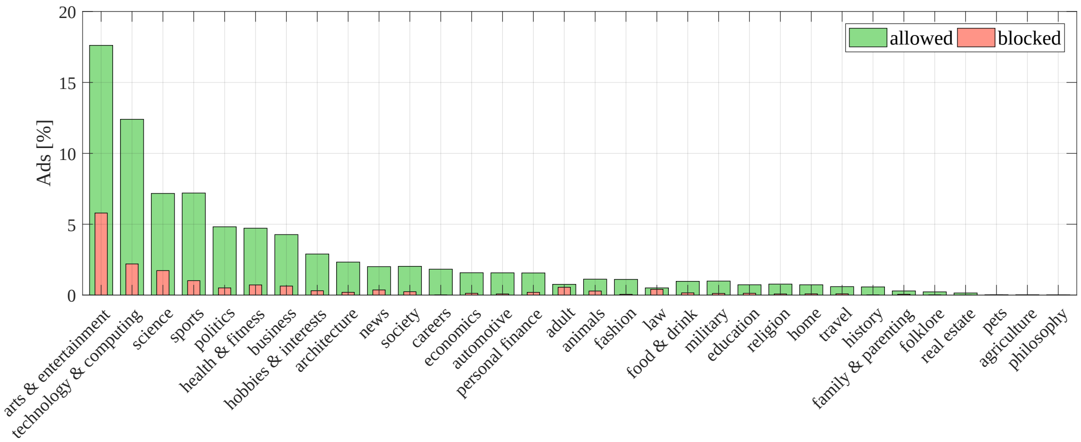

Fig. 4: Distribution of ads per category and depending on whether such ads were blocked or allowed.

with the top 30 ad domains that delivered ads through iframes. An eye-opening result is that nearly 40 percent of all ads shown to our 746 users were served by two Google's domains, namely, doubleclick.net and googlesyndication.com.

Fig. 4 illustrates the percentage of ads per page category that were blocked or allowed according to the anti-tracking policies specified by our users. The categories of web pages hosting most of the ads were "arts & entertainment", "technology & computing" and "sports". Not entirely unexpected, we find that the ads most affected by blocking were those displayed in the categories most blocked by users, namely, "adult", "arts & entertainment" and "health & fitness", with a ratio of blocked ads to total ads of 42.13, 24.74 and 17.42, respectively. However, since the number of pages belonging to such categories was comparatively less visited by our users (and such pages had comparatively lesser number of ads), the total number of ads blocked by all users of MyTrackingChoices was rather small. Overall, as few as 16.25% of all ads were blocked by our tool, which accounts for a significant reduction in the effect of ad-blocking, when compared to the existing technologies that eliminate 100% of ads.

Having examined the overall percentage of blocked ads, next we aim to estimate the ultimate economic impact of said ad-blocking on the Web. To this end, we capitalize on the analysis conducted by previous own research [42], where we estimated the prices paid by different types of advertisers in 2014 to display their ads through RTB. To the best of our knowledge, ours is the only study on ad-price estimation per topic category for desktop browsing. For the sake of completeness, we provide the list of prices in Table III, where the average cost per thousand impressions (CPM) and the standard deviation (SD) are shown for 11 interest categories.

Table III: Average prices per category in CPM

Category	Avg. price	Std
Adult	0.25	0.15
Humor	0.25	0.19
Sports	0.29	0.18
Games	0.32	0.16
Blogs / Web Communications	0.33	0.25
Entertainment	0.33	0.23
Computers / Internet	0.38	0.24
News / Media	0.38	0.26
Society / Lifestyle	0.38	0.27
Vehicles	0.41	0.34
Reference	0.48	0.21
Restaurants / Food	0.59	0.31
Shopping	0.68	0.38

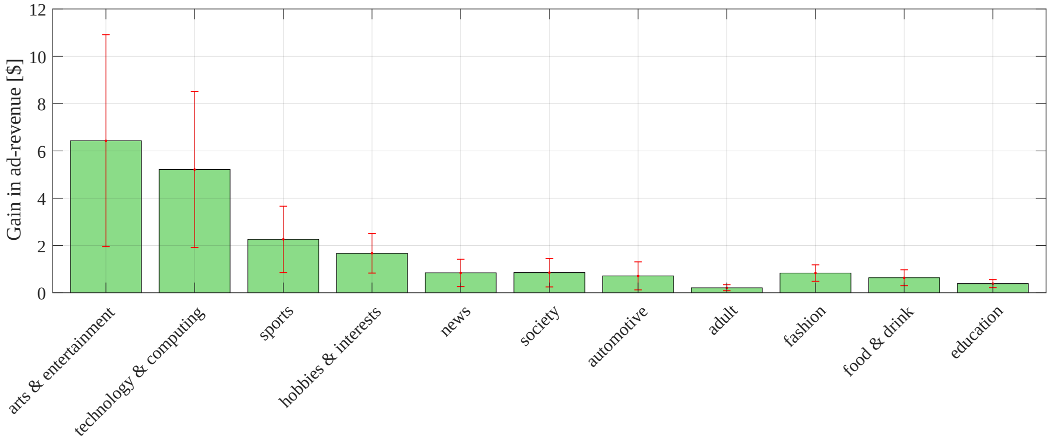

Fig. 5: Estimated gain in ad-revenue in our data set. One SD above and below the mean value (green) is depicted.

The idea is to estimate the impact of our tool based on such prices and our results of ad blocking per category. However, we must take into account that RTB is one of the two third-party advertising models (see Sec. 2) and that the money paid by advertisers through traditional ad platforms might differ substantially from RTB auctions. In the absence of any information about the ad-prices paid through these ad platforms and owing to the fact that RTB is the dominant technology, we assume the same estimates given by [42].

We illustrate the estimated impact of fine-grained control over tracking in Figs. 5 and 6. In the former figure, we plot the amount of money that publishers —and ad platforms through a small fee charged per served ad— "gained" as a result of 746 users using MyTrackingChoices from April 2016 to April 2017. By *gain*, we mean that said ad revenue would have been lost if all our users had chosen an ad blocker like Adblock Plus to protect their privacy. We compute this gain based of the average price per category and the total number of ads which were not blocked as a result of a blocking policy. We would like to emphasize that because such prices are not available for all our 32 categories (as shown in Table III), our results are restricted accordingly. In Fig. 6, on the other hand, we represent the loss in ad-revenue due to our anti-tracking technology.

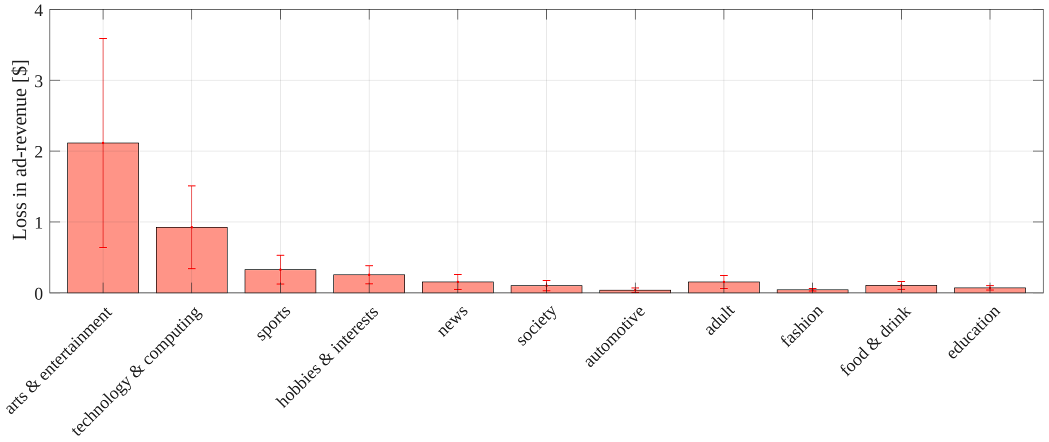

Fig. 6: Percentage of ads blocked and allowed per page category. One SD above and below the mean value (red) is depicted.

The most remarkable result that follows from these figures are the reduced gains in ad-revenue brought by MyTrackingChoices for each category[15]. However, it must be noted that this is simply a consequence of the low CPM prices reported in [42], which average 0.00039 dollars per ad, and also the moderate number of impressions displayed to users of our data set, namely 110 642 served in 1 624 142 pages. The reason for this ratio of ads to browsed pages may be found, in turn, in the data-collection module integrated into MyTrackingChoices, which can only capture static-image-based ads (which are typically iframe-based) and thus it is unaware of any textual, Flash-based and video ads shown to our users. It is important to stress, nevertheless, that a substantial portion of such browsed pages might in fact correspond to search engines and e-mail services, where typically no display ads are present. But because it is not possible for us to ascertain the actual domain names of the hashed URLs, we cannot find out the actual portion. Hence, in the absence of any rigorous study on the number of ads displayed to desktop browsing users, and as a way to validate our data-collection procedure, we can only mention that Alexa's 500 top sites contains numerous web pages which are free of ads. As a matter of fact, among the top 50 most popular sites on the Web, only 56% of them currently show ads for desktop users on their home pages.

In addition to this reduced overall impact by the users of our data set, we also notice that the publishers providing "arts & entertainment" and "technology & computing" and "sports" content were the ones most affected, positively and negatively, by our plug-in.

Concluding remarks. Trying to extrapolate the above results to a larger population of users adhered to MyTrackingChoices is a challenging task, and so is obtaining a reliable estimate of the overall impact of our tool on the Web economy. For privacy reasons, hardly any demographic information of our users is available to us, and thus it is difficult to assess whether our data set is representative of the whole population of users or not. From the Firefox version of our plug-in, however, we know that most users are from the U.S., France and Germany, and that all users (Chrome and Firefox) in our data set downloaded the plug-in directly from the corresponding Web repositories.

[15] In our estimation of those gains we disregarded the 32.77 percent of the allowed ads for which there was no price available, according to Table III.

In the absence of such information, we can only assume that our set of users might resemble, to a certain extent, the whole population. This is obviously under the assumption that users would be motivated for installing an add-on which, in addition to preventing tracking more granularly, might support content providers. Bearing this in mind, our result that 48.13% of the pages browsed by our 746 users were blocked by MyTrackingChoices might be an indicator that the cost of ad blocking could be dropped by nearly half.

However, ad blocking has several dimensions and does not only affect desktop browsing but also mobile browsing and mobile apps. This means that the most reliable estimate of the loss of global revenue due to blocked advertising [15], which was 21.8 billion dollars in 2015, may not be suitable for our purposes. Besides, while in that same year mobile was starting to get into the ad blocking game, mobile ad-block usage has now surpassed desktop ad blockers according to a recent report [23].

The problem with said estimation of the cost of ad-blocking based on the percentage of blocked ads is that, first, the number of ads need not be uniformly distributed across all websites; and secondly, ad-pricing depends on the particular page ads are delivered, among other aspects. For these two reasons, our study also examined the number of blocked ads in our data set on the basis of the publisher type. Our analysis, although it is limited to iframe-based ads and relies on the ad-prices reported for the RTB model, indicated a greater reduction in the impact of ad blocking when compared to that estimated from the blocked pages. In particular, we observed that just 16.25 percent of all ads were blocked by our tool.

Although the overall monetary impact was estimated to be just 20.06 ± 5.91 (SD) dollars in ad-revenue, it must be noted that this was essentially due to the low ad-prices reported for desktop browsing [42]. In fact, ad prices for mobile browsing have been recently estimated to be in the range of \$0.01 CPM to \$100 CPM, with an average of \$25 CPM [43]. This in contrast with the average of \$0.39 CPM for desktop browsing used in our analysis, which suggests that the potential gain in ad-revenue provided to publishers by a massive adoption of our tool (by both desktop and mobile users) is here underestimated.

Last but not least, it is important to bear in mind that our conclusions about the magnitude of the impact of MyTrackingChoices on the Web economy are inextricably dependent on the reliability of the CPM estimates. The data we rely on is, to the best of our knowledge, the only source information with ad prices for desktop browsing on a per-category basis. The more recent study cited above ([43]), albeit for mobile advertising, shows even larger deviations in the estimation of this pricing information. For example, the authors of this study reported CPMs of 0.3 ± 3 (SD) dollars approximately for ads related to "technology & computing", whereas we have considered CPMs of 0.38 ± 0.24 (SD) dollars from [42] for that same category.

To reduce this uncertainty in the estimation of MyTrackingChoices' impact on the Web economy —an uncertainty due to the unavailability of reliable estimates of ad prices—, we indicate in Sec. 7.1 some possible improvements and open research directions.

5. USABILITY AND PERFORMANCE EVALUATION

In this section, we examine some aspects related to usability and performance. We report upon the pseudo-anonymous data collected from our users.

5.1. Usability

Usability is an essential factor for a wide scale adoption of any Internet tool. In our case, the measurement of this factor is a challenging task since no direct user-feedback in this regard is available to us. In this section, we analyze two aspects of our tool that

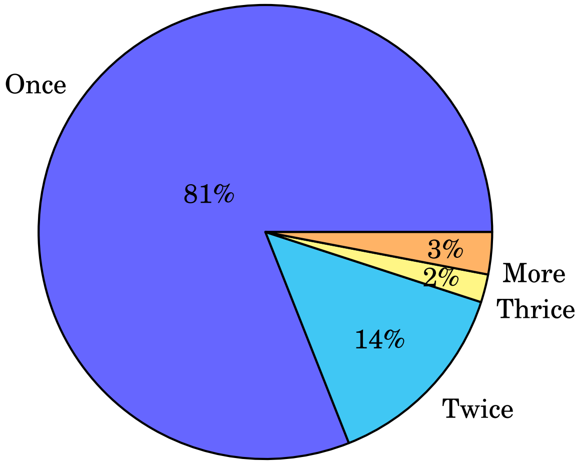

Fig. 7: Percentage of users modifying the category-based anti-tracking policy different number of times.

may be related to its usability. However, bear in mind that, in the absence of said feedback, the analysis is more oriented to providing some indicative figures about the usage of MyTrackingChoices, rather than attempting to evaluate its usability.

Before going into further details, recall that current anti-trackers allow users to block ads based on a per URL basis, whereas our tool lets users specify their blocking policies based on categories of web pages. That said, in an attempt to evaluate the usability of our tool we compute the frequency with which users needed to change the blocking policies. Under this first criterion, we compare the blocking policies used by current anti-trackers and MyTrackingChoices. It is evident that category-based blocking should outperform URL-based blocking. However, for completeness and accuracy, we measure the performance of MyTrackingChoices in this regard.

We note that, in certain circumstances, category-based blocking can be too coarse. For example, a user may not want to block a specific web page in a blocked category or vice versa. To serve this purpose, MyTrackingChoices offers an additional possibility to be more granular and block or allow on a per web page basis. Our second criterion captures the usage of this feature. This criterion is evaluated by counting the number of users who exercised such a granular feature and at how many web pages this feature was used.

Next, we measure the afore-described criteria on our data set. First, we study how users' tracking choices evolve over time. Specifically, we want to know if users are satisfied by just configuring once their tracking choices when they install the tool or they change their tracking choices post-installation.

Fig. 7 shows the percentage of users who changed their anti-tracking policies. We note here that the vast majority of our users defined said policies once and then did not feel the need to modify them. This suggests that the proposed category-based blocking is stable over time and that our tool does not require much tuning once such preferences are clear. Additionally, we observed that almost all remaining users added one or more categories, while only 9 users removed an already blocked category.

Having seen the frequency with which users change their tracking preferences, now we examine whether and how often users prevented tracking through the URL-blocking option. We find that 239 users in our data set utilized this functionality on 0.71% of the browsed pages. Although this is not a considerable percentage, this result indicates that such an option is a necessary and complementary strategy to category-

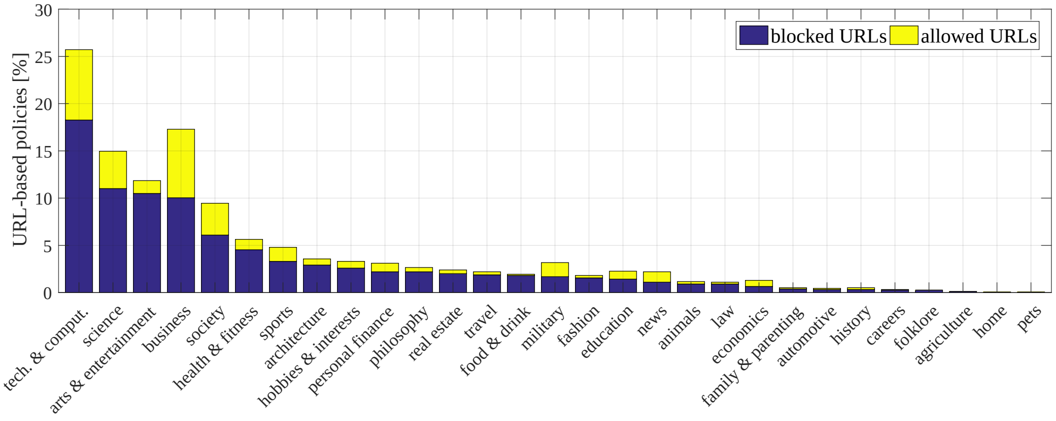

Fig. 8: Distribution of per URL defined policies by category.

based blocking. The reason is that, through this more granular control, users can exert control over tracking whenever they are not satisfied with the plug-in's per category-blocking decision.

Fig. 8 shows the distribution of per-URL-granular policies specified by those 239 users over different categories of web pages. It is worth noting that, among the top-5 categories for which per URL policies are defined, none of them belonged to the top-5 categories blocked by users (Fig. 2). This means that the web pages which contributed the most to the per-URL policies are those which correspond to non-sensitive categories. Also, we observe that 34.45% of the per-URL policies were defined to allow trackers on different web pages where tracking was blocked as per their category. This leads us to the conclusion that some users are willing to accept ads on web pages belonging to non-sensitive category and indeed want to support publishers. However, at the same time, we also notice that trackers/ads are blocked on a per-URL basis for web pages which are not blocked based on their categories. This implies that users are disposed to block trackers and thus eventually ads on certain web pages, even if they did not consider their categories sensitive. Such a user behavior suggests that these web pages might have contained intrusive ads and for this reason users opted for blocking third-party domains.

5.2. Performance

Our tool aims to return control to users over third-party tracking by enforcing their privacy preferences. However, the control returned to users comes at the cost of some processing and computational overhead. In this section, we evaluate this cost.

We measure the performance of MyTrackingChoices in various respects. First, we measure the performance of our tool in terms of correctness of categorization algorithm and proper functioning of the web pages. We find that incorrect categorization was reported for only 31 web pages (out of a total of 1 624 142 distinct web pages browsed by all our 746 users) by 24 users. This seems to indicate that users are mostly satisfied with the categorization results provided by our tool, which is in line with our evaluation of the page-classification algorithm conducted in Sec. 3.2.1. Also, 12 users reported 12 web pages whose functionality broke due to MyTrackingChoices. Most of the times it was because of blocking of content-provider domains and we corrected the problem by updating the list of allowed domains.

Next, we assess our extension in terms of average page-loading time, maximum memory usage and maximum CPU usage, and compare these results with some of

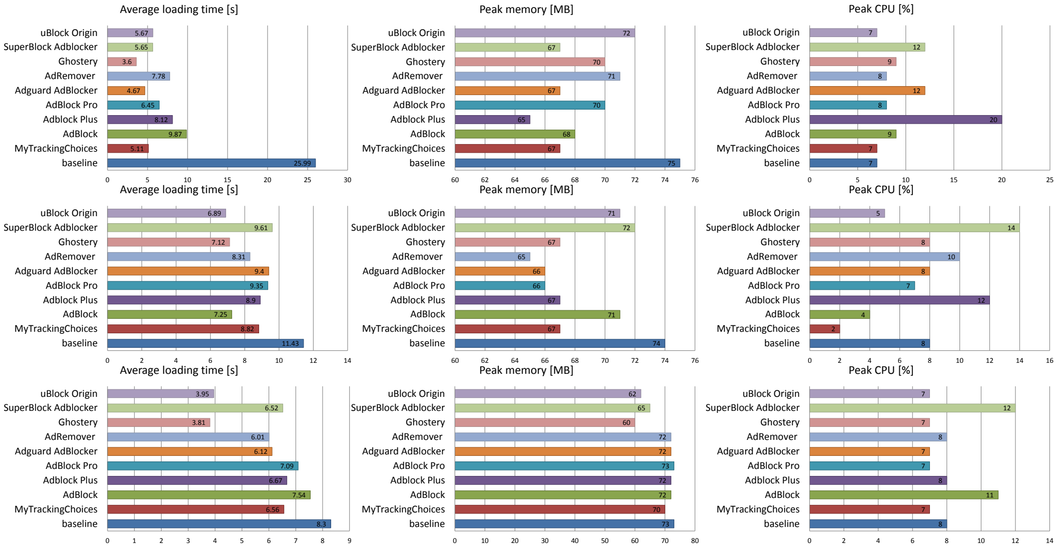

Fig. 9: The first row shows the average loading time (in seconds), peak memory usage (in MegaBytes) and peak CPU usage (in %) for the home page of CNN website (cnn.com). The second and third rows show the same information for the home pages of Le Monde (lemonde.fr) and New York Times (nytimes.com) respectively.

the most prominent ad-blockers and anti-trackers. This is necessary because our extension provides certain privacy guarantees in terms of user tracking but it comes at the cost of processing and computational overhead.

Fig. 9 presents the benchmark analysis of this comparison as well as a baseline scenario where no ad-blocking tools are used. The results of this benchmark analysis have been obtained after ten consecutive visits to these three popular web pages: cnn.com, lemonde.fr and nytimes.com. Obviously, any caching mechanism was prevented in our evaluation. The main conclusion that can be drawn from this figure is that our technologies, despite its higher complexity, performed reasonably well, compared with most of the tools evaluated in this study. In terms of page loading time, MyTrackingChoices seems to be close to Ghostery and uBlock Origin (except for the site nytimes.com), the two tools which led this aspect. As for memory usage, our plug-in required between 67 and 70 MB, ranking second —together with SuperBlock Adblocker and Adguard AdBlocker— after Adblock Plus in cnn.com, and ranking fourth after Ghostery, uBlock Origin and SuperBlock Adblocker in nytimes.com. Finally, with a maximum processing usage between 2 and 7 percent, our extension outperformed all tested tools in terms of CPU consumption. In a nutshell, the performance of MyTrackingChoices was observed to be about the average for those three web pages, and this was despite the inclusion of more sophisticated and advanced privacy features.

6. PRESENCE OF TRACKERS

In this section, we study the prevalence of trackers on the Web. We first discuss how this study is conducted and then present the measurement results.

If a third-party domain is present on a web page, we assume that it is tracking users. Moreover, a third-party domain is classified as “tracker” only if it is found to be tracking users on three or more web pages, as mentioned in Sec. 2. The consequence is that the first two presences of a third-party domain does not suffice for the domain to be classified as a “tracker”. In addition, we exclude from our study those web pages which

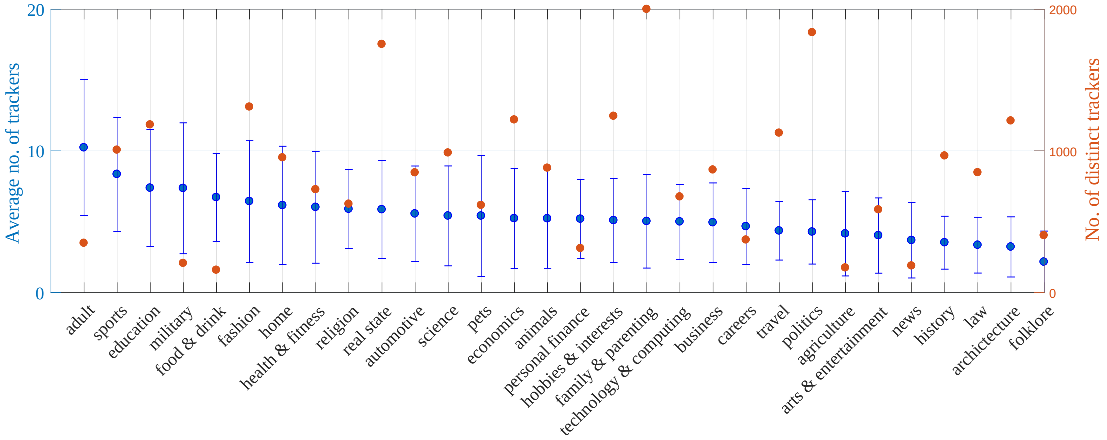

Fig. 10: Distribution of trackers per content category. The left-side y-axis represents the average number of trackers with SD, whereas the right-side y-axis provides the number of distinct trackers.

are blocked by users themselves. This is because if the initial request to a tracker is blocked, none of the consequent requests, which would have been made otherwise, can be captured.

Users in our data set were tracked 2 554 592, times by a total of 2 781 distinct trackers. The average number of trackers over all users per page is 3.02, with SD of 5.16. However, the maximum number of trackers present on a web page is as high as 151 in some cases whereas the minimum is zero.

In Fig. 10, we plot the distribution of trackers across different categories. From this figure, we note that those pages classified as “home” contain the greatest number of trackers on average. However, if we contrast it with the number of ads received on those pages (Cf. Fig. 4), we find that the web pages in the “home” category did not receive the greatest number of ads. This seems to indicate that web pages of such category include many additional trackers other than those related to ads. In fact, the average number of trackers in said category is 10.22 and the SD is 4.79. This is much higher than the global average of 3.02 trackers per page.

From the same figure, it is interesting to note that the “technology & computing” pages exhibited the largest number distinct trackers, close to 2 000, even though the average number of trackers per web page is quite small. Actually, those web pages categorized as “home” are tracked by 349 different trackers, although the average number of trackers per page is the highest one. This result might suggest that trackers are most spread in web pages belonging to the “technology & computing” category.

Finally, we would like to stress that more extensive experimental research on Web tracking [28] appears to be in line with our results, particularly with those concerning with the prevalence of tracking per interest category.

7. CONCLUDING REMARKS

In the last few years, as a result of the proliferation of intrusive and invasive ads, the use of adblocking and anti-tracking tools have become widespread. The problem with these technologies is that they pose a binary choice to users and thus disregard the crucial role of advertising as the major sustainer of the Internet’s free content.

We believe that such technologies are only a short-term solution, and that better tools are necessary to solve this problem in the long term. Most users are not against ads and are actually willing to accept some ads to help Web sites [15; 14; 11; 17; 21].

However, this is provided that, in this ad-delivery process, users can control the personal information gathered.

We have proposed MyTrackingChoices, a Web technology that may address this situation by giving users real control over their browsing histories. The proposed tool empowers users to take control and reinforce their tracking preferences, and its ultimate aim is to strike a better balance between user privacy and the Web economy.

The proposed tool aims at protecting users' privacy by selectively regulating the information disclosed to ad companies and Internet trackers. However, at the same time, it allows users to enable tracking on those web pages which are not considered to be sensitive, which is crucial for preserving the Internet business model whereby users get free content and services in exchange of receiving ads.

Among other features, our technology facilitates a simple albeit effective management of user privacy. Through a user-friendly interface, users can select which of the 32 top-level categories they do not want to tracked on. While browsing the Web, they are informed about the category of the visited pages as well as the trackers available on such pages, raising awareness about the ubiquity of tracking and the entities following their traces. Also, the fact that our tool may block tracking on sensitive categories is in compliance with current regulations on data protection.

To estimate the impact that the proposed tool would have on the Internet, we have conducted a thorough analysis with real-world user data. Among other results, our analysis shows that the 48.13% of the pages browsed by our set of users and the 16.25% of the ads displayed to them were blocked by MyTrackingChoices. Based on the latter figure, which suggests an important reduction in the impact of ad blocking, as well as the ad-prices reported by previous own work, we estimated small benefits for publishers as a result of our more granular ad blocking. However, these benefits to the online advertising ecosystem are underestimated since our analysis is restricted to static-image-based ads and desktop-browsing, and the ad-prices are not available for all our publisher categories.

7.1. Limitations and Future Work

We conclude this section by highlighting the limitations of our analysis of the impact of MyTrackingChoices on the Web economy. Next, we identify some strands of future research.

One of the main limitations of our analysis is that the current version of our tool is only able to detect iframe-based ads. Since these type of ads are mostly delivered in the form of images (either static or dynamic), our study is not aware of any textual or video ads displayed to our users. Besides, our analysis is constrained by the fact that MyTrackingChoices is available only to desktop browsing on Chrome and Firefox, which means that the impact of fine-grained control over mobile browsing and advertising is not addressed.

On the other hand, our estimate of the economic impact due to MyTrackingChoices is affected by our non-perfect page classifier, which shows a good performance for sensitive categories such as "adult", "health & fitness" and "religion" (95.98%, 91.15% and 91.02% of correctly classified pages, respectively), but is less accurate for other popular content categories like "news" (77.77%) and "travel" (80.12%).

Furthermore, we are also restricted by the scarce data on CPM prices. We build on the results of previous own research [42], which is the only study on ad-price estimation per topic category for desktop browsing. However, said data is not available for all our publisher categories and more importantly exhibits large deviations, which causes uncertainty in the estimation of the economic impact of our tool.

To cope with such limitations and further support our arguments on the suitability of fine-grained control over per-page tracking, next we explore possible improvements and open research directions.

The Web page categorization algorithm currently uses the landing page, i.e., the page where the user is redirected to when they click on the image-based ad, to categorize the page. As the landing page is not always available, the image itself could be used, in addition to or along with, the landing page information. There already exists much off-the-shelf software for optical character recognition in images[16].

The other important direction of future work is to reduce the uncertainty in the estimation of the economic impact of MyTrackingChoices as a result of the large deviations in CPM prices reported by previous work. This is needed to convince users and also all other stakeholders, i.e., advertisers and publishers, that our approach could actually be useful to simultaneously protect privacy and provide free content to users. We intend to diminish said uncertainty by examining the clear-text prices for winning bids which are often included in the request and response of RTB messages. Because these data are available on the user side, our plug-in should be able to capture them. With this information, we could remove the effects that the ad-serving time, browsing profiles, devices and location may have on CPMs. Because our tool would have access to the prices of the ads delivered to our users, we could compute a better estimate of the actual cost of our tool on the Web economy.

On the other hand, since privacy is only one of the reasons users block ads, MyTrackingChoices could be extended to provide fine-grained choices on the other main reason behind ad blocking, namely, intrusiveness. Specifically, MyTrackingChoices could ask users for the maximum number and types of ads they would like to see per page, and then enforce their choices technically. We believe this could help our tool gain more adoption since the addition of this feature will not only attract the privacy conscious users but also those concerned with intrusive ads

Acknowledgment

We would also like to thank the anonymous reviewers for their immensely helpful suggestions to improve the readability and contents of this article The writing quality of this paper is benefited by comments from Amrit Kumar. This work is partly funded by the Inria Project Lab CAPPRIS, by the European Commission through the H2020 project “CLARUS”, and by the Spanish Ministry of Economy and Competitiveness (MINECO) through the project “Sec-MCloud”, ref. TIN2016-80250-R. J. Parra-Arnau is the recipient of a Juan de la Cierva postdoctoral fellowship, FJCI-2014-19703, from the MINECO.

REFERENCES

[1] Adblock plus. https://adblockplus.org. accessed on 2016-02-22.

[2] Disconnect. https://disconnect.me/. accessed on 2016-02-22.

[3] Ghostery. https://www.ghostery.com. accessed on 2016-02-22.

[4] Iab tech lab publisher ad blocking primer. Technical report. accessed on 2016-03-13.

[5] Privacy badger. https://www.eff.org/fr/privacybadger. accessed on 2016-02-22.

[6] Rip: Adblock plus. http://www.engadget.com/2016/02/12/rip-adblock-plus/.

[7] YourOnlineChoices. http://www.youronlinechoices.com/. [Online; accessed 17-February-2016].

[8] Handbook on European data protection law. http://www.echr.coe.int/Documents/Handbook_data_protection_ENG.pdf, 2014. [Online; accessed 17-February-2016].

[9] A Way to Peace in the Adblock War. https://blogs.harvard.edu/vrm/2015/09/21/a-way-to-peace-in-the-adblock-war/, 2015. [Online; accessed 21-February-2016].

[16] For example, https://github.com/deborausujono/crfocr.

[10] Acceptable Ads Manifesto. https://acceptableads.org/, 2015. [Online; accessed 17-February-2016].

[11] DCN Consumer Ad Block Report. https://digitalcontentnext.org/blog/press/digital-content-next-research-indicates-33-of-consumers-likely-to-try-ad-blocking-software-in-next-three-months/, 2015. [Online; accessed 17-February-2016].

[12] Firefox interest dashboard. https://github.com/mozilla/interest-dashboard/blob/master/refData/IAB.js, Nov. 2015. [Online; accessed 21-February-2016].

[13] Getting LEAN with Digital Ad UX. http://www.iab.com/news/lean/, 2015. [Online; accessed 17-February-2016].

[14] Profiling Adblockers. http://www.globalwebindex.net/blog/profiling-adblockers, 2015. [Online; accessed 17-February-2016].

[15] The cost of Ad Blocking, PageFaire and Adobe Ad Blocking Report. http://downloads.pagefair.com/reports/2015_report-the_cost_of_ad_blocking.pdf, 2015. [Online; accessed 17-February-2016].

[16] Tracking preference expression (DNT). Technical report, Aug. 2015.

[17] Why Millennials Block Ads. http://www.dmnews.com/opinions/why-millennials-block-ads/article/475448/, 2015. [Online; accessed 17-February-2016].

[18] Brave. https://brave.com/, 2016. [Online; accessed 17-February-2016].

[19] Efficient & optional filtering of domains in Recursor 4.0.0. http://blog.powerdns.com/2016/01/19/efficient-optional-filtering-of-domains-in-recursor-4-0-0/, 2016. [Online; accessed 21-February-2016].

[20] List of allowed third-party domains. https://myrealonlinechoices.inrialpes.fr/allowed_thirdparty_domains.json, 2016. [Online; accessed 17-February-2016].

[21] Why do people block online ads? https://www.adspeed.com/Blog/people-block-online-ads-1844.html, 2016. [Online; accessed 17-February-2016].

[22] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *Proc. PostCOLING Workshop Multiling. Ling. Resources*, pages 101–108, Hangzhou, China, Aug. 2004.

[23] S. Blanchfield. Adblocking goes mobile. Res. rep., PageFair, May 2016.

[24] F. Chanchary and S. Chiasson. User perceptions of sharing, advertising, and tracking. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 53–67, 2015.

[25] R. Cookson. Google, Microsoft and Amazon pay to get around ad blocking tool. *Financial Times*, Feb. 2015. accessed on 2016-02-22.

[26] Corpus of global web-based english.

[27] J. Daudé, , L. Padró, and G. Rigau. Validation and tuning of wordnet mapping techniques. *Proc. Int. Conf. Recent Adv. Nat. Lang. Process. (RANLP)*, Sept. 2003.

[28] S. Englehardt and A. Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proc. ACM Conf. Comput., Commun. Secur. (CCS)*, pages 1388–1401. ACM, 2016.

[29] T. Gerbet, A. Kumar, and C. Lauradoux. A Privacy Analysis of Google and Yandex Safe Browsing. Research Report RR-8686, INRIA, Feb. 2015.

[30] S. Goel, J. M. Hofman, and M. I. Sirer. Who does what on the web: A large-scale study of browsing behavior. In *ICWSM*, 2012.

[31] A. Gonzalez-Agirre, E. Laparra, and G. Rigau. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proc. Global WordNet Conf.*, 2012.

[32] M. Gundlach. AdBlock. https://getadblock.com/. accessed on 2016-02-22.

[33] A. Kae, K. Kan, V. K. Narayanan, and D. Yankov. Categorization of display ads using image and landing page features. In *Proc. ICDM Workshop Large-Scale Data Min.: Theory, Appl.*, pages 1–8. ACM, 2011.

[34] J. B. Kristen Purcell and L. Rainie. Survey on search engine use. http://www.pewinternet.org/2012/03/09/main-findings-11/. accessed on 2016-02-22.

[35] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 561–570, New York, NY, USA, 2010. ACM.

[36] P. G. Leon, B. Ur, Y. Wang, M. Sleeper, R. Balebako, R. Shay, L. Bauer, M. Christodorescu, and L. F. Cranor. What matters to users?: Factors that affect users' willingness to share information with online advertisers. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 7:1–7:12, New York, NY, USA, 2013. ACM.

[37] B. Magnini and G. Cavaglià. Integrating subject field codes into wordnet. In *Proc. Lang. Resource, Evaluation (LREC)*, pages 1413–1418, June 2000.

[38] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[39] D. Marti. Targeted Advertising Considered Harmful. http://zgp.org/targeted-advertising-considered-harmful/. accessed on 2016-02-22.

[40] W. Melicher, M. Sharif, J. Tan, L. Bauer, M. Christodorescu, and P. G. Leon. (Do Not) track me sometimes: Users contextual preferences for web tracking. In *Proc. Int. Symp. Priv. Enhanc. Technol. (PETS)*, Lecture Notes Comput. Sci. (LNCS), pages 1–20. Springer-Verlag, 2016.

[41] G. A. Miller. WordNet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[42] L. Olejnik, T. Minh-Dung, and C. Castelluccia. Selling off privacy at auction. In *Proc. Symp. Netw. Distrib. Syst. Secur. (SNDSS)*. Internet. Soc., Feb. 2014.

[43] P. Papadopoulos, N. Kourtellis, P. R. Rodriguez, and N. Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay for your personal data? In *arXiv: 1701.07058v2*, Mar. 2017.

[44] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[45] J. Turow, J. King, C. J. Hoofnagle, A. Bleakley, and M. Hennessy. Americans reject tailored advertising and three activities that enable it (september 29, 2009). http://ssrn.com/abstract=1478214 or http://dx.doi.org/10.2139/ssrn.1478214. accessed on 2016-02-22.

[46] S. Whitbeck. RTB is growing like mad. is your mobile marketing keeping up? Technical report, 2015.