The Online Tracking Horde: A View from Passive Measurements

Hassan Metwalley, Stefano Traverso, Marco Mellia, Stanislav Miskovic, Mario Baldi

The Online Tracking Horde: a View from Passive Measurements*

Hassan Metwalley[1], Stefano Traverso[1], Marco Mellia[1], Stanislav Miskovic[2], and Mario Baldi[1 2]

[1] Politecnico di Torino, Italy
{metwalley,traverso,mellia}@tlc.polito.it
[2] Symantec Corp., USA
{stanislav_miskovic,mario_baldi}@symantec.com

Abstract. During the visit to any website, the average internaut may face scripts that upload personal information to so called online trackers, invisible third party services that collect information about users and profile them. This is no news, and many works in the past tried to measure the extensiveness of this phenomenon. All of them ran active measurement campaigns via crawlers. In this paper, we observe the phenomenon from a passive angle, to naturally factor the diversity of the Internet and of its users. We analyze a large dataset of passively collected traffic summaries to observe how pervasive online tracking is. We see more than 400 tracking services being contacted by unaware users, of which the top 100 are regularly reached by more than 50% of Internauts, with top three that are practically impossible to escape. Worse, more than 80% of users gets in touch the first tracker within 1 second after starting navigating. And we see a lot of websites that hosts hundreds of tracking services. Conversely, those popular web extensions that may improve personal protection, e.g., DoNotTrackMe, are actually installed by a handful of users (3.5%). The resulting picture witnesses how pervasive the phenomenon is, and calls for an increase of the sensibility of people, researchers and regulators toward privacy in the Internet.

1 Introduction

Internet is the revolution that changed our life, allowing us to be informed, buy goods, enjoy shows, play games, keep in touch with friends, and freely express our opinions to potentially very large audiences. People are more and more connected to the Internet, with mobile terminals allowing access to information from anywhere, anytime. Companies see the Internet as a means to stay in contact with their customers, to attract them, and to offer more and more personalized content. Not surprisingly, a large fraction of Internet businesses rely on online advertising, a market that keeps growing year by year, and that generated $42B revenue in 2013 according to the Interactive Advertising Bureau [1].

Online advertisement – ads for short – enables companies to design very targeted campaigns. The web offers the capability of reaching specific groups of users assembled with very fine granularity, leveraging knowledge of personal interests and taste

* This work was conducted under the Narus Fellow Research Program.

of individuals. In order to collect such knowledge, companies track the users during their everyday online activity, constantly collecting information for marketing purposes (e.g., products browsed on a shopping website, online newspapers usually read, movies liked). This information is used to profile a user in order to deliver tailored ads, recommend movies to watch, or goods to buy.

Online trackers play a key role in this ecosystem as third-party services that "shadow" users during their browsing activity. Trackers rely on host of solutions to identify a user, ranging from storing a cookie on the user browser or device to exotic tracking techniques that fingerprint users across several web sites [7,14,18]. The tracker business models also vary greatly. Some offer customized ads, while others sells user information to ads companies, acting as data brokers. Google's DoubleClick and Yahoo's YieldManager are notable tracker examples. However, the full list of companies that build their business around information collection includes several hundreds.

The mechanisms associated to tracking users can be beneficial for both companies and consumers. But they also raise many privacy concerns among the regulators and researchers. Ultimately, the consciousness of the people about their privacy being violated in the Internet is growing day by day.

Several works in the literature study the latest advances in online tracking, unveiling new and more subtle mechanisms [7,13,16,17], and proposing countermeasures to be protect users'privacy, typically in the form of browser plugins [2,3,4,5,6]. Some works studied the pervasiveness of online tracking by running active measurement campaigns and by crawling the web [15,8,12]. This paper falls in this second class: we aim at quantifying the pervasiveness and extensiveness of online tracking. Differently to any previous work, we are the first, to the best of our knowledge, to leverage passive measurements, which have the major advantage to naturally factor the users into the picture. We address questions as how many tracking services an internaut would normally face during her activity? How different is the picture from past years, or from different vantage points? How invasive are tracking services?

For our study, we rely on an extensive dataset composed by (anonymized) traces we collect by passively observing normal users from four different probes installed in two ISPs in two different countries. We use this data to pinpoint the traffic exchanged with a list of online trackers that we manually built from various sources. We then collect statistics to characterize such traffic.

Results confirm what is known from the literature: online tracking is ubiquitous. We count more than 400 active online tracking services, with 100 of them being regularly contacted by more than 50% of users, and the most pervasive ones that are impossible to avoid. Results confirm observations shown in other works based on active campaigns [16,12], but our passive approach naturally factors the user browsing behavior, and allows us to obtain a very detailed and fine-grained picture that quantifies the pervasiveness of tracking services in real life. For instance, 77% of users face the first tracker just 1 s after starting their online activity. We observe websites that nowadays embed more than 50-100 third-party trackers, attracted by the chance to monetize visits, and in practice contributing to collect personal information. Notably, most of these services are not popular to enter in the top list of websites (and thus have never being considered by active studies). Yet, those are popular enough to collect a sizable number of

users. Our unique vantage point allows us to measure things that active campaigns can not gauge. We are the first at quantifying the popularity of privacy-enhancer browser plugins. Surprisingly this is limited, with DoNotTrackMe installed by a mere 3.5% of users. This testifies the small consciousness and sensibility of internauts versus their privacy. Similarly, by splitting statistics by type of user device, we highlight how Android devices are more prone to interact with tracking services than iOS devices and regular Windows PCs. Finally, our measurements highlight another phenomenon: the increase adoption of HTTPS as the means to collect data. This exacerbate the tension on the need to protect users privacy, since for instance this mines the possibility to develop in-network solutions to control and limit online tracking services.

We hope the picture we draw can contribute to increase the sensibility of people, researchers and regulators towards privacy in the Internet. We do not believe in an "arms race" as a possible solution, but rather in a solution in which people is offered the means to take informed choices.

The remainder of the paper is structured as follows. Sec. 2 introduces the related work, Sec. 3 details the dataset we employ in this study, Sec. 4 presents the results, and, finally, Sec. 5 concludes the paper.

2 Related Work

Our work is related to recent literature in the area of web measurement driven studies about web tracking and online advertisement. We can divide most of the notable works in this area in three branches. The first branch is mostly oriented to understand which identifiers and techniques online tracking services exploit to record users' browsing activities. Yen et al. [18] examine the common identifiers trackers can leverage to identifying users, and the authors of [16] and [11] describe the techniques third party trackers and online social networks use to monitor the activity of their users. The second branch is mostly oriented to understand the leakage of personal information due to web trackers. For instance, Balachander et al. [15] studies privacy leakage and evolution of third party trackers over four years from 2005 to 2008 using DNS logs. Another notable example is [14], which analyzes how popular websites share users' private information with tracking services, remarking that this trend is worryingly diffused. The last branch focuses about the analysis of the mechanisms which drive online advertisement. Vallina et al. [17] specifically examine ads in mobile terminals. Our aim is different as we address the problem of understanding how pervasive tracking services are by leveraging a large set of passive measurements. And to the best of our knowledge, we are the first to perform this analysis following a passive approach.

Our study shares some common points with other works. We discuss in the following the differences which distinguish our work. Barford et al. [8] build their analysis around a dataset they collect thanks to a web crawler they develop. Web crawling lets the authors infer detailed information about the online ads which populate webpages. They specifically focus on the analysis of online ads, while our study addresses online tracking services in general, and from users' perspective.

The work which mostly approaches this study is Gomer et al. [13], where authors propose a methodology to identify tracking services from the analysis of pages returned

Trace	Probe	Period	IP addr	Services
ISP1-Vp1-1d-05/12	ISP1-Vp1	09/05/2012	11660	200320
ISP1-Vp1-1d-05/13	ISP1-Vp1	08/05/2013	12218	239230
ISP1-Vp1-1d-05/14	ISP1-Vp1	07/05/2014	10458	238617
ISP1-Vp1-1d-02/14	ISP1-Vp1	26/02/2014	11027	247797
ISP1-Vp2-1d-02/14	ISP1-Vp2	26/02/2014	11927	297488
ISP2-Vp1-1d-02/14	ISP2-Vp1	26/02/2014	4911	113648

(a) One-day long traces.

Trace	Probe	Period	IP addr	Services
ISP1-Vp1-10d-10/14	ISP1-Vp1	13-23/10/2014	13408	1046339
ISP1-Vp2-10d-10/14	ISP1-Vp2	13-23/10/2014	11149	1306612
ISP1-Vp3-10d-10/14	ISP1-Vp3	13-23/10/2014	1321	415550

(b) Ten-day long traces.

Table 1. The sets of traces we consider in this study.

by search queries. Specifically, using Google, Bing and Baidu, they run popular queries (extracted from the 2005 KDD dataset). They then crawl the top 10 returned pages, and check for the presence of trackers embedded in each page.

Similarly, [10] and [12] offer a global point of view of this phenomenon. In first case, Castelluccia et al. [10] analyze the provenance of most important third party tracking services using two popular browser extensions, AdBlock Plus and Ghostery, for the geographical classification. In second case, Falahrastegar et al. [12] crawl the top websites in Alexa rank for different countries, and measure the per-country pervasiveness of third party trackers. Despite our study share the same aim, as said, we rely on passive traces. This allows us to naturally factor the interactions with third-party trackers and real internauts during their daily activities, and check the impact of multiple devices, or browsers, or even malware eventually being installed on end-users' terminals.

Finally, we are the first to analyze in a real scenario the adoption of those do-not-tracking extensions as AdBlock Plus or DoNotTrackMe which are expected to protect users' privacy. Thanks to our vantage point, we show the breadth of the most invasive trackers, considering both the services hosting them, and the users' chances to contact them.

3 Dataset

In this work, we employ four different passive probes running Tstat [3] that we installed in Points-of-Presences (PoPs) in the operational networks of two different ISPs (ISP1 and ISP2) in Europe. Tstat observes all packets flowing on the links connecting the PoP to the ISP backbone network. It rebuilds each TCP flow, tracks it, and at the end of the flow, logs more than 100 detailed statistics in a simple text format. For instance, for each TCP flow, Tstat logs the anonymized client IP address, the server IP address, the application (L7) protocol type, L7-bytes sent and received, etc. Tstat also implements DN-Hunter, an advanced mechanism that allows to annotate each TCP flow with the server hostname the client resolved via DNS before actually contacting the server IP address [9]. For TCP connections carrying HTTP and HTTPS data, DN-Hunter has been proved to unveil the *service* being contacted, e.g., `www.acme.com` or `mail.acme.com`. For HTTP traffic, Tstat produces a separate log which details most relevant HTTP fields: the HTTP method (GET, POST, etc.), the server hostname, the URL path, the referer, the client user-agent, etc.[4]

[3] `tstat.polito.it`

[4] The traffic logs Tstat generates do not contain information which may offend ISP users privacy. Indeed, Tstat processes IP packets (and their payload) in real time and generates transport- and

We have been collecting TCP and HTTP logs since May 2012. In this work, we focus on a subset of the data. Specifically, we consider three probes (ISP1-Vp1, ISP1-Vp2 and ISP1-Vp3) that are located in PoPs of the same ISP (ISP1), in two different cities of the same country. A fourth probe (ISP2-Vp1) is installed in a different ISP (ISP2), in a second country. Tab. 1 describes, for each trace, the name used throughout the paper, the location, the period, the number of households (identified by the IP address of the access gateway, see next section), and the total number of different services, i.e., server hostnames. Tab. 1(a) refers to traces which are one-day long, from the same probe, collected on the second Wednesday of May 2012, 2013 and 2014. We complement them with three traces we collected in the same day (February 26th, 2014), but from three different PoPs and countries. Tab. 1(b) refers to ten-day long traces that we collected from three different PoPs of ISP1, during the same period in October 2014.

Among the details, Tab. 1 shows that the dataset we use covers several thousands of regular users, which browse some millions of hostnames. Results we present are as such generic, even if specific to the country where the vantage points are located.

3.1 Identifying Active Users and Number of Connected Devices

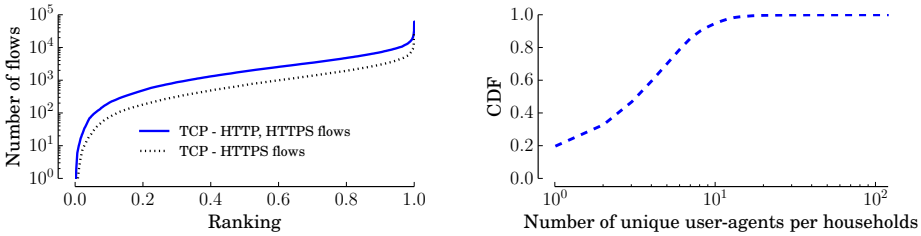
Notice that the client IP address field in our logs refer to the access gateway (ADSL/FTTH modem) customers are given by the ISP. As such, the IP address can be considered as an identifier of the household, which may hide several actual devices and users that connect to the Internet using NAT at the access gateway. This includes possible households in which no actual user is present, but in which some device generates some traffic. In particular, ISP1 offers native VoIP communications. The access gateway acts also as VoIP gateway, thus we expect some households to appear as "active" (IP address is used) even if no terminal is present (we observe VoIP data only).

Fig. 1(a) shows the number of total HTTP and HTTPS flows per each IP address, sorted for increasing number. Notice the log scale on y-axis. Plot refers to trace ISP1-Vp1-1d-05/14. It shows the bias induced in ISP1 by the presence of VoIP gateways at the modem. They indeed generate some signalling HTTP and HTTPS traffic to report VoIP usage statistics to the operator. The presence of the sharp knee in the figure suggests that a simple threshold based filter is sufficient to identify "active" households.[5] In the remainder of the paper, we take a conservative choice, by considering active only those IP addresses for which we see at least one HTTPS flow, and at least 100 HTTP or HTTPS total flows for the 1-day long dataset (1000 flows for the 10-day long traces). This filters out those sources of traffic we are not interested in (e.g., smart TVs, VoIP gateways, or pure P2P clients). Most active households reach $100,000$ flows per day.

To quantify the presence of multiple clients that are hidden behind the NAT at the ADSL router, Fig. 1(b) shows the CDF of the number of different user-agents seen for

HTTP-level logs in which we take care of obfuscating any privacy sensitive information (e.g., IP addresses are anonymised using irreversible hashing functions, all URLs are truncated, etc.). Second, Tstat has no visibility on encrypted traffic (HTTPS), where the sensitive information concentrates. Furthermore, we report that our traffic monitoring activity is approved by the Security Office of the ISP in which we deploy our probes.

[5] We use the term "household" and "user" interchangeably in the paper.



(a) Number of flows per household. Trace ISP1-Vp1-1d-05/14.

(b) CDF of user-agents seen behind a single household. Trace ISP1-Vp1-10d-10/14

Fig. 1. Per-household statistics.

a given active IP address. We consider only user-agents associated to actual browsers, for PC and mobile terminals. We leverage the `User-agents 0.3.1` Python library for this.[6] Results show that only 20% of households have only one terminal, with 75% of them showing between 2 to 10 different user-agents. Manually checking this, we observe a lot of smartphones and tablets, with some cases showing multiple browsers being normally used. Surprisingly, in few cases we see more than 10 user-agents. A manual check shows the presence of suspicious behavior with possibly a malware generating lots of HTTP requests toward few IP addresses serving advertisements. HTTP requests contain a rotating set of legitimate browsers user-agents. We suspect this to be related to some click fraud activity, i.e., a malicious user artificially generating clicks on ads servers by forging user-agents. We noted the presence of these outliers, and check their presence is not affecting our statistics in the remainder of the paper.

3.2 Identifying Online Tracking Services

We build a list of online tracking services by merging together data we obtain from different sources. First, inspired by the approach used by the authors of [12], we instrument a browser to visit the top 500 websites of the global Alexa rank. For each page, we visit it and use the Ghostery plugin [2] to pinpoint the presence of trackers. Given then the hostname of the tracker server, we extract only the second-level domain name to reduce the list. For instance from `cnt2.acmetracksyou.com` and `srv1.acmetracksyou.com` we consider `acmetracksyou` only. We repeat the procedure using the top 500 websites in the general Alexa rank, and of countries where ISP1 and ISP2 are located. By merging the resulting lists we obtain more than 350 distinct online tracker services. Then, we complement this list with the one obtained from the developers of Abine[7], and with some specific trackers we manually identify. This list also includes hostnames referring to trackers specifically tailored to track mobile clients. The final list consists of 443 distinct online tracking services. The list includes only services that we classify as third party sites that collect users' information, and eventually serve advertisements. This

[6] `https://pypi.python.org/pypi/user-agents`
[7] `https://www.abine.com/`

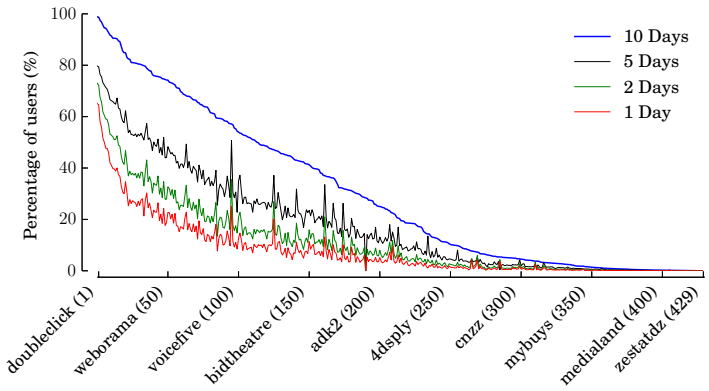


Fig. 2. Penetration of online trackers in ISP1-Vp1-10d-10/14.

includes tracking services that profile users explicitly (e.g., Doubleclick) or that track users when on a website (e.g., Google Analytics). We do not consider social network buttons, plugins, and active code.

In the remainder of the paper, we rely on this list to pinpoint connections that clients establish with tracker servers. When analyzing the TCP logs, we use the DN-Hunter hostname to identify traffic to trackers. For HTTP logs instead we use the server hostname in the HTTP request.

4 Results

4.1 Penetration of Online Tracking Services

We start our analysis by measuring the "penetration" of each online tracker that appears in our list. We consider the trace ISP1-Vp1-10d-10/14. Fig. 2 reports the percentage of users that contacted at least one time a given tracker with respect to users that we find active considering the entire 10 days. The results are shocking: the top online tracking services – DoubleClick, Google Analytics, and Google Syndication – track 98.8%, 98.7% and 97.4% of users, i.e., as soon as a user goes online, sooner or later she/he will contact one of Google tracking services. While they might be known to some users, the list of trackers includes a vast majority of players that are mostly unknown even to experts internauts. Fig. 2 reports some of the names. Observe the solid blue curve which refer to the ten-day long period. More than 50% of users contacts 120 distinct trackers, with 429 out of 443 trackers that have being contacted by at least one user. When considering shorter periods of time, e.g., one, two, five days (red, green and black curves), the number of users seen by trackers decreases. Yet, the top 20 trackers can observe more than 35% of internauts active during the first day of the trace.

Next, we compare penetration of trackers over years. Fig. 3 shows the results. This time we are considering one-day long traces during October 2012, 2013 and 2014, and we focus on the top 23 trackers. Penetration is higher in this case as we compute it

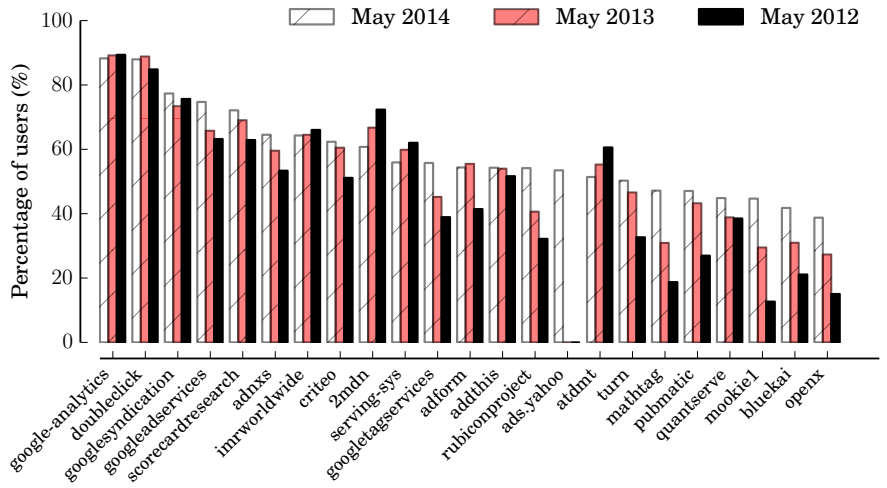


Fig. 3. Penetration of online trackers compared for different years in ISP1-Vp1-1d-05/*.

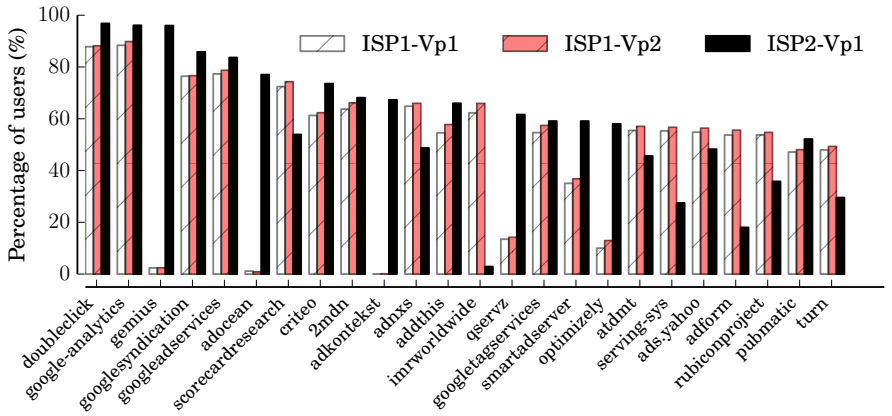


Fig. 4. Penetration of online trackers compared for different years in ISP1-Vp1-1d-02/14, ISP1-Vp2-1d-02/14 and ISP2-Vp1-1d-02/14.

over the active population of a single day. Top trackers show marginal changes over year, reflecting the fact that they have saturated the coverage. Going down in the list, we see that most of trackers shows an increase in the penetration, with only few exceptions. Some new players shows up, i.e., `ads.yahoo`. No service went out of business (or disappeared).

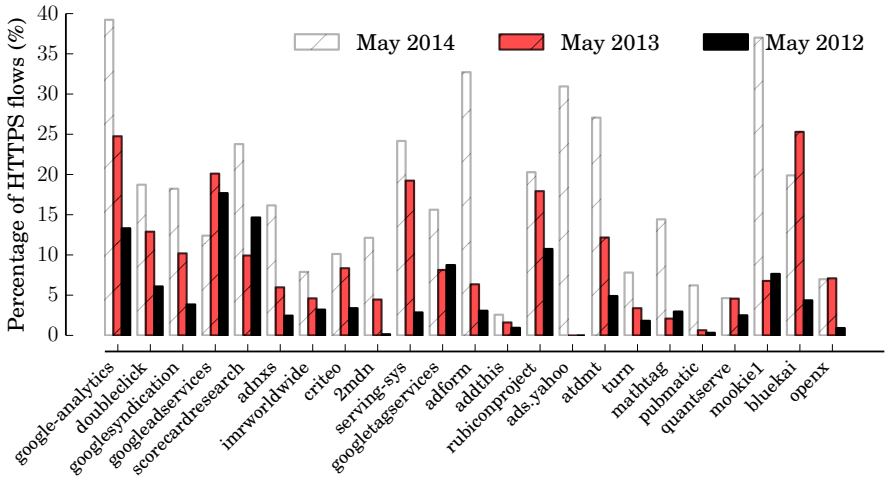


Fig. 5. Percentages of HTTPS transactions users establish with third party tracking services for different years in ISP1-Vp1-1d-02/1*.

We next compare the penetration from different vantage points. Fig. 4 show results considering the top-30 most popular trackers seen in the merged list of ISP1 and ISP2. Again, one-day long traces (in February 2014) are considered. Two observation holds: ISP1-Vp1 and ISP1-Vp2 show practically the same results. Despite being in two different cities, the population interest and habits is very similar, being in the same country. Conversely, comparing ISP1 and ISP2 results, we observe a very different penetration for some trackers, which reflects a localized service. This confirms the finding in [12] which highlighted the different coverage of online tracking services. For instance, Google tracking services present a higher penetration in Country 2 than in Country 1, while some trackers do not cover ISP2 market, e.g., `imrworldwide`, and viceversa, e.g., `adocean`.

To gauge the amount of data trackers collect, we compute the distribution of the fraction of TCP flows to trackers (not reported for the sake of brevity). We observe that 60% of users exchange from 10% to 30% of flows with trackers. We also see few cases in which more than 95% of flows are sent to trackers. Investigating, we observe i) click fraud activity of some users infected by some ad-malware, and ii) some mobile application that keeps downloading tens of ad-banners per minute, for hours. Both are likely illicit behaviors caused by malicious attackers that abuse of unaware users to game the ads market.

At last, we measure to which extent the top trackers rely on encrypted channels, i.e., HTTPS, to collect information about the users. To this end, we measure how many TCP flows the users exchange with the trackers, and how many of these flows are HTTPS. We consider again the one-day long traces we collected in 2012, 2013 and 2014. Fig. 5 reports the percentages of connections carrying HTTPS traffic for the top 23 most popular trackers. We observe some trackers do use encryption to collect users' information: in

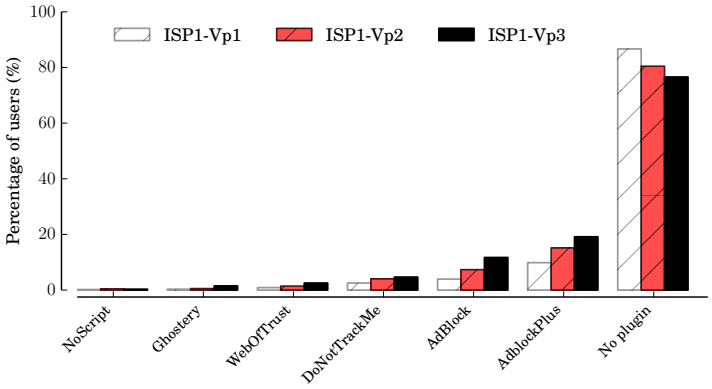


Fig. 6. Percentage of households installing the most popular do-not-tracking plugins. Results from traces ISP1-Vp*-10d-10/14.

2014, *google-analytics* (39%), *adform* (33%), *ads.yahoo* (31%) and *mookie1* (36%). In general almost all the top 23 tracker has consistently increased the usage of HTTPS over the last three years. This is also mandated by the general increase of HTTPS-enabled websites that enforce HTTPS for all third party content too.

4.2 Popularity of Privacy Enhancer Plugins

We now investigate the popularity of plugins that can enhance and customize the browsing experience. We focus on those well-know plugins which i) block Javascript code commonly used by advertisers (NoScript), ii) warn about the presence of online tracking service (Ghostery, WordOfTrust, DoNotTrackMe), and iii) block advertisement traffic (AdBlock, AdBlockPlus). We count how many users run these plugins. For each plugin, we perform some active experiments to understand which hostnames it has to contact to check if updates are available[8] We then compute the fraction of users that contact such hostnames in our traces.

We report in Fig. 6 the shares of households that have installed a given plugin in at least one device, together with the percentage of those which did not install any plugin in any device. We compute these statistics for all our ten-day long traces. As shown, the share of users installing a plugin is in general rather small, and it seems that users are more interested in blocking the ads they encounter while browsing the web, rather than trackers. Indeed the popularity of AdBlockPlus is between 10% to 18%. Less than 3.5% of users run DoNotTrackMe. Moreover, we observe that more than 80% of the users do not install any of the considered plugins, thus offering the trackers the capability of easily following their surfing activity.

[8] We observe that each plugin contacts its update server with a fairly large frequency (e.g., at any browser bootstrap or once a day) with respect to the considered observation window.

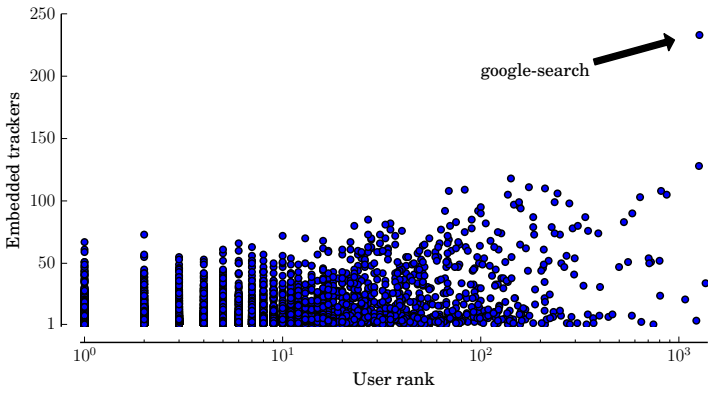


Fig. 7. Scatter plot of the number of users contacting a service, and the number of trackers embedded by the same service. ISP1-Vp1-1d-5/14.

4.3 Trackers Penetration among Services

Next, we investigate the penetration of the online trackers among different services (e.g., websites) that users contact during their everyday online activity. To this end, we consider the HTTP trace. From each URL where the `hostname` is a given tracker, we check the `Referer` field to observe which service was embedding it. As before, we consider only the second level domain name as the name of the service. We count more than 25,000 services that host third party trackers. For each of them, we count how many users contacted them, i.e., how popular they are, and how many and which trackers they embed. In the scatter plot in Fig. 7, each black dot represents a service; the x-axis (in log scale) reports the number of distinct users accessing it; the y-axis reports the number of embedded trackers. Data refers to ISP1-Vp1-1d-5/14. The scenario is rather heterogeneous, with many services embedding several tens of tracking services. We observe both unpopular services hosting many trackers – e.g., the few services contacted by one or two users only, but hosting more than 50 trackers – and popular services hosting a few trackers – e.g., the rightmost bottom corner of the plot. In general the number of trackers per service tends to increase with the popularity of the service.

When checking the results we noticed that `www.google.com`, the most popular service, apparently embeds 222 trackers. By manually digging in our traces we observe that such large number of trackers is due to a bug in the *google-search* widget installed in version 4 of Android devices which affect our (and possible other datasets). Indeed, when an Android user performs a search query using the widget, and then visit a webpage by tapping on a link in the returned page, the Android browser keeps using `http://www.google.com` as `Referer` field for all objects that actually refer to the visited page. Besides affecting our results, this bug can possibly also poison the counters that trackers use to share their revenues with services.

To complement the above observations, we analyze tracker's "breadth", i.e., the number of services that embed each tracker. We report the results in Fig. 8. The x-axis reports the rank of the 150 most popular trackers in our lists, and the y-axis reports

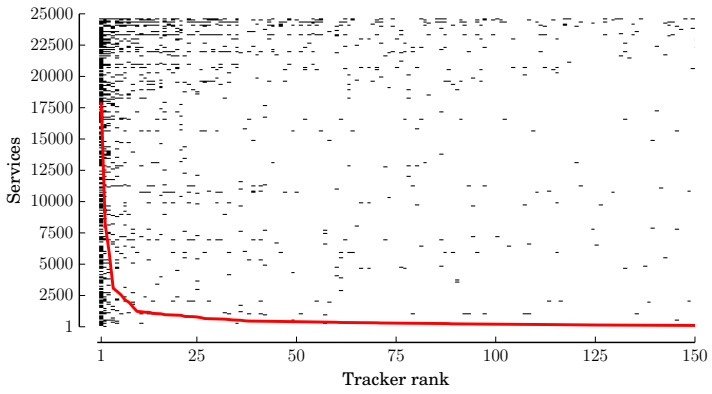


Fig. 8. Trackers embedded by the services users visited (black dots), and the number of services covered by each tracker (red curve). trace ISP1-Vp1-1d-5/14.

all the services present in ISP1-Vp1-1d-05/14. Each dot represents the association of tracker x with the service y. We sort the services by considering their popularity among the users, from the most popular (top) to the least popular (bottom). First, most popular services embeds a large variety of trackers (observe the dense area in the top part of the plot). This confirms the trend of Fig. 7. Second, the dense vertical area in the leftmost part of the plot indicates that trackers with the highest penetration are also associated to many services (and vice-versa). To ease the visualization, the red solid curve shows the number of services associated to each tracker in the rank. The curve is very steep, with less than 10 trackers are associated to more than 1000 services. In particular, the three trackers with the largest service coverage belong to Google: *google-analytics*, *doubleclick* and *googlesyndication*, embedded by $17,814$, $8,176$, and $5,921$ services, corresponding to 71%, 32%, and 24% of the total number of active services, respectively. The first tracker not belonging to Google in the rank, *addthis* takes the fourth place with $3,080$ (12%) covered services. Despite this, it sees more than 50% of population (see Fig. 3). This reflect a market dominated by Google, in which a lot of other small players are present. The ones that are hosted in popular domains are able to still track a lot of users.

4.4 Time to Be Tracked

In this section, we investigate how invasive trackers are at getting in touch with users. We measure how much time a user spends online before encountering the first tracker. Let $T0$ be the time of the first HTTP or HTTPS TCP flow generated by a user, and let $T1$ be the time of the first TCP flow to a tracker. We measure the Time-To-Tracker as $TTT = T1 - T0$. For this analysis, we leverage the TCP trace in ISP1-Vp1-1d-5/14. We consider only those households for which we know just PC-based terminals are used, i.e., only one single PC-based user-agent is seen. Results are astonishing: TTT is smaller than 1 s in 77% of PC users, i.e., as soon as a users goes online, she/he hits the

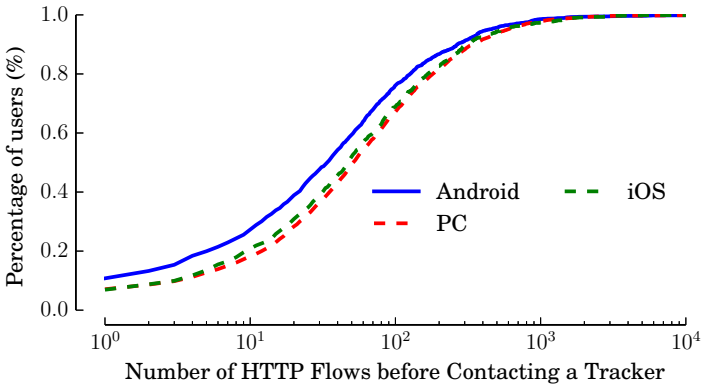


Fig. 9. CDF of number of HTTP requests before contacting a tracker. Trace ISP1-Vp1-1d-5/14.

first tracker in less than a second. Even worse, 100% of users have a TTT smaller than 100 s. When considering all households, we observe even shorter TTT.

To give the intuition behind this, we detail the number of HTTP connections needed to hit a tracker using the HTTP traces. We split the dataset according to the user-agent field in three categories: PC-based, Android-based and iOS-based. For each category, we compute the distribution of the number of HTTP requests a user generates before contacting a tracker. Fig. 9 plots the results. Independently on the device, in about 10% of the cases, the first HTTP transaction of the day goes to a tracker[9], and within the first 100 (1000) requests in about 60% (97%) of the cases. Interestingly, users with an Android device contact a tracker earlier on than users with a PC or an iOS-based device.

5 Conclusions

Motivated by the privacy concerns that online tracking services have recently raised, we presented in this paper a passive characterization of this phenomenon. We leveraged a large dataset of traffic summaries we collected from ISPs located in two different countries to passively quantify the pervasiveness and the intrusiveness of online tracking services in our life. To the best of our knowledge, we are the first to passively analyze the behavior of trackers in a real scenario, where the users are naturally factored.

The results presented in this paper are boggling. We observed that top 100 trackers collect information from 50% of the users on a regular basis. Plus, some of these being able of tracking 98% of the Internauts. They are embedded into more than 70% of websites, including the most popular ones, but also those that are visited by few users. Similarly, trackers' intrusiviness is astonishing, with 77% of users that contacts a tracker within 1 second after she/he starts browsing the web. We also observe that trackers are increasingly embracing HTTPS to collect data. While this is possibly driven by the

[9] The reason why the first HTTP request goes to a tracker is due to the user browsing on HTTPS before moving to HTTP, thus becoming visible for this measure.

increase of HTTPS usage, this increases the warning level since is becomes more and more complicated to control and limit the information they can collect.

Our results show that the consciousness of the users about their activity being monitored by trackers is limited. Indeed, only a small fraction of users rely on privacy-enhancer browser plugins as DoNotTrackMe, and they appear to be more interested in ad-blocking extensions such as AdBlockPlus.

We believe that the information contained in this paper can contribute to increase the consciousness of people about the fragility of their privacy in modern web. We hope that our findings may be of stimulus for regulators, researchers and practitioners who aim at designing solutions to let the users take control of the information they exchange with the Internet.

References

1. IAB internet advertising revenue report, 2013 full year results, http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2013.pdf
2. Ghostery, https://www.ghostery.com/en/
3. DoNotTrackMe, http://www.abine.com/donottrackme.html
4. Privacy Badger, https://www.eff.org/privacybadger
5. AdBlockPlus, http://adblockplus.org/
6. WoT, https://www.mywot.com/
7. Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., Diaz, C.: The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In: ACM SIGSAC. (2014)
8. Barford, P., Canadi, I., Krushevskaja, D., Ma, Q., Muthukrishnan, S.: Adscape: Harvesting and Analyzing Online Display Ads. In: WWW. (2014)
9. Bermudez, I.N., Mellia, M., Munafo, M.M., Keralapura, R., Nucci, A.: DNS to the Rescue: Discerning Content and Services in a Tangled Web. In: ACM IMC. (2012)
10. Castelluccia, C., Grumbach, S., Olejnik, L.: Data Harvesting 2.0: from the Visible to the Invisible Web. In: WEIS. (2013)
11. Chaabane, A., Kaafar, M.A., Boreli, R.: Big Friend is Watching You: Analyzing Online Social Networks Tracking Capabilities. In: ACM WOSN. (2012)
12. Falahrastegar, M., Haddadi, H., Uhlig, S., Mortier, R.: The Rise of Panopticons: Examining Region-Specific Third-Party Web Tracking. In: TMA. (2014)
13. Gomer, R., Mendes Rodrigues, E., Milic-Frayling, N., Schraefel, M.: Network analysis of third party tracking: User exposure to tracking cookies through search. In: ACM WI-IAT. (2013)
14. Krishnamurthy, B., Naryshkin, K., Wills, C.E.: Privacy leakage vs. Protection measures: the growing disconnect. In: W2SP. (2011)
15. Krishnamurthy, B., Wills, C.: Privacy Diffusion on the Web: A Longitudinal Perspective. In: WWW. (2009)
16. Roesner, F., Kohno, T., Wetherall, D.: Detecting and Defending Against Third-party Tracking on the Web. In: USENIX NSDI. (2012)
17. Vallina-Rodriguez, N., Shah, J., Finamore, A., Grunenberger, Y., Papagiannaki, K., Haddadi, H., Crowcroft, J.: Breaking for Commercials: Characterizing Mobile Advertising. In: ACM IMC. (2012)
18. Yen, T.F., Xie, Y., Yu, F., Yu, R.P., Abadi, M.: Host Fingerprinting and Tracking on the Web: Privacy and Security Implications. In: NDSS. (2012)