When Sally Met Trackers: Web Tracking From the Users' Perspective

Savino Dambra[1]
Eurecom
Norton Research Group

Iskander Sanchez-Rola[1]
Norton Research Group

Leyla Bilge
Norton Research Group

Davide Balzarotti
Eurecom

Abstract

Web tracking has evolved to become a norm on the Internet. As a matter of fact, the web tracking market has grown to raise billions of dollars. Privacy cautious web practitioners and researchers extensively studied the phenomenon proving how widespread this practice is, and providing effective solutions to give users the option of feeling private while freely surfing the web. However, because all those studies looked at this trend only from the trackers' perspective, still there are a lot of unknowns regarding what the real impact of tracking is on real users. Our goal with this paper is to fill this gap in the web tracking topic. Thanks to logs of web browsing telemetry, we were able to look at this trend from the users' eyes. Precisely, we measure how fast a user encounters trackers and research on options to reduce her privacy risk. Moreover, we also estimate the fraction of browsing histories that are known by trackers and discuss two tracking strategies to increase the existing knowledge about users.

1 Introduction

Third-party web tracking was first introduced to support web analytics and advertisement [30] but evolved over the years into a very widespread phenomenon employed for a wide range of purposes. Currently, more than 90% of the websites include at least one tracking script [16, 50], resulting in a multi-billion dollar business [21, 29, 36, 66] where many companies earn huge amounts of money by selling or leveraging the data collected from users.

Previous works showed that users are aware of this practice and have rightfully started to complain about the amounts of online tracking present on the web [46, 64]. On the other hand, those studies also reported that participants are surprised when confronted with detailed information about the extent and prevalence of web tracking [34, 64]: once aware of the actual impact, users' general attitudes often resulted in being at odds with such practices [34], and in stronger intentions to take privacy-protective actions [64].

The scarcity of works that investigate how impactful web-tracking is for Internet users can explain why, despite being aware of the practice, only a few are conscious of the actual implications and take the appropriate actions to protect themselves. For example, only 7.74% of the browsers' market share belongs to privacy-centered browsers [26], 8.5% of the users reported the use of tracker-blocking tools [64], and just 0.59% of them use privacy-preserving search engines [54]. We believe that studies that look at the problem from the users' perspective to identify concrete evidence for its seriousness could be immensely helpful to the general population.

In fact, as mentioned earlier, web tracking is not a new phenomenon on the Internet and a wide corpus of previous works have analyzed both the impact and the prevalence of web tracking. However, previous studies have assessed its size by measuring how many websites contain trackers, or how many websites are known to a given tracking company [8, 24, 38, 55]. As we will demonstrate in this work, knowing in how many websites a tracker is detected is difficult to translate into how much the tracker knows about the average user. More than that, our experiments show that measuring the coverage by only crawling top-ranked websites results in gross under-estimation. In reality, users visit only a tiny fraction of the Internet websites – typically composed of a mix of popular (such as social networks, search engines, news) and less popular sites (such as regional pages, friends' blogs, or specific work-related sources). As a result, it is still unknown what fraction of the user's browsing history is known to web trackers or what fraction of trackers are encountered by each user.

In this work, we aim at filling this gap by complementing the current knowledge on web tracking with real-user browsing behaviors. We leverage the telemetry of 250K users and the information collected by a large-scale crawling experiment to analyze the impact that web tracking has on end-users located all around the world. Differently from previous studies, whose results are based on the analysis of the top websites listed on publicly available services [3, 4], the use of browsing telemetry allows us to exactly know when and which websites are accessed by users, without the need for distribution approximations. This allows us to precisely understand how often users

[1] Savino Dambra and Iskander Sanchez-Rola contributed equally to this work as first authors

encounter new trackers, how many different ones, and what amount of information each tracker knows about them. As privacy advocates, we were extremely careful to preserve the privacy of the users in our dataset. All of our data was anonymized, and raw browsing histories were processed in an automated fashion, presented and analyzed only in aggregated form.

The paper is organized in two parts. First, we look at the web-tracking from a time and frequency perspective: for each user in our dataset, we estimate how long it takes to encounter all and a significant fraction of the trackers. We then perform a correlation analysis to understand what increases the privacy risk, discovering that there is an interesting relationship among privacy and security risks on the web. In the second part of the study, we estimate what percentage of the user's browsing history is known to trackers and investigate how much this knowledge could be extended through real or hypothetical collaborations among different tracking companies. For instance, our experiments show that the actual knowledge popular trackers have of the users' histories is almost **double** the estimate obtained by crawling the top Alexa popular domains. We also shed light on the most efficient monitoring strategy and what sensitive information could be learned about the users because they browse particular classes of websites.

We hope that our findings could bring awareness to the users and motivate them to use privacy-preserving solutions to prevent web tracking.

2 Background And Related Work

The first tracker, based on a cookie from `digital.net` in `microsoft.com`, was used in 1996 and discovered by an 'archaeological' study conducted by Lerner et al. [30] in 2016 by using the Internet Archives Wayback Machine [23].

The first analysis regarding web tracking was performed in 2009 by Krishnamurthy and Wills [27], where they examined the different technical ways in which third-parties could obtain user-related information. Three years later, the work from Mayer and Mitchell [33], and Roesner et al. [45], helped to lay the foundations for future studies. More recent studies showed that an increasingly larger percentage of the most popular websites include some form of tracking, and that they use a variety of techniques to do it [16, 24, 50, 55].

Olejnik et al. [40] were among the first to use real-user data to study web tracking. The authors discovered that 69% of the users in their dataset had a fingerprint that could differentiate them from the rest based on their web history. This study was recently replicated by Bird et al. [10] with 52K Firefox users, and found an even larger number, with 99% of them showing unique patterns. Falahrastegar et al. [19] also used the web history of real users to check whether user-specific IDs were being sent in requests: authors found this to be very common between certain groups of domains. Vallina et al. [61] performed instead a study based on network traffic of a mobile carrier to check not only the presence, but also the efficiency of the ecosystem based on energy consumption. They found that tracking is very widespread but the delivery strategy is inefficient.

During the last years, the number of works based on real-user data has increased. In 2018, Karaj et al. [25] performed a large-scale study using the information gathered from a browser extension. They calculated some general stats about the different trackers found online, and open-sourced the corresponding global results obtained from the dataset. At the same time, Papadopoulos et al. [42] presented a study focused on mobile devices. By using the data collected from 1,270 users, the authors quantified the economical cost of showing ads for companies, and the corresponding privacy loss by the users that receive them. The final results indicate that there is a clear imbalance between the two, with the users paying the highest price. The following year, Papadopoulos et al. [41] expanded their idea and analyzed the concept of tracking cookie synchronization by using another dataset of 850 real mobile users. They found that 97% of the users are actually exposed to this type of practices in the first week of browsing. Most recently, the work from Hu et al. [20] leveraged real-world browsing histories to measure the prevalence of different tracking organizations in UK and China. Authors discovered that there is a big difference in the companies involved, with home-grown third-party operators in China, and US players dominating the UK market. Finally, Mishra et al. [35] studied the relevance of the IP information in the web tracking ecosystem, analyzing the information received from 2,230 users. Results indicate that IP-based tracking is still a viable, as 87% of the participant retained the same address for multiple days.

In summary, many papers analyzed web tracking by using different types of telemetry, but they centered their work on very specific cases such as user identifiers [19, 41] or web history uniqueness [10, 40, 61]. Despite finding many interesting results, these studies lack a global overview of: i) the perspective of how the user arrives to that tracking situation, and ii) what is the strategy and knowledge that trackers follow. In this work, we try to find answers to these two questions.

3 Data Sources and Methodology

Our main dataset comes from the telemetry of a popular security company. The data, collected on the consumer hosts about the users' *web-browsing activity* is described in Section 3.1. We acquire the *category and risk score* (Section 3.3) for each domain in the telemetry and detect the *trackers* present on the webpages by using a custom crawler (Section 3.2). We also take advantage of a *linkage graph* published by Sanchez-Rola et al. [49] about the information-sharing relationships among different trackers (Section 3.4).

Each piece of information— from its collection, throughout its analysis, to its storing— is treated in a way that preserves the customers' privacy and identity. We never deanonymize users by looking at their browsing sessions and we only look at aggregated data. The authors had multiple discussions

(before and during the study) with the legal department of the company to get the approval for this study and make sure that the data was processed ethically and preserved the users' anonymity. In this respect, we detail all the adopted measures for each of the datasets in its dedicated subsection.

3.1 Web-browsing telemetry

This dataset contains the web-browsing history of 250K users. The telemetry is collected by the company's antivirus (AV) sensor installed on Windows machines and only includes users who *voluntarily* install the product, accept the company's privacy policy [39], and opt-in to share their data. The user identifier is anonymized on the client-side and sent in this form to a central system: in our study, we observe users only through numeric anonymized identifiers, that do not contain any detail or endpoint attribute able to trace back to their origin. The telemetry spans a period of 8 days and was collected from October 14th to 21st of 2019. The data includes a code that reports the country registered by the user when installing the AV software, a daily log with the list of domains browsed by each user, and the hour in which the request was performed. Overall, we count 2.35M distinct websites (0.8% were not accessible or offline), which finally accounted for 107M entries in the users' browsing history.

3.2 Website trackers

We identify the trackers that exist on the websites in our dataset through a custom crawling framework. Note that, to further preserve the privacy of the users, the necessary tracking-relation information is collected without any human intervention and nothing else related to the content of the website is collected. The framework has been developed in the first months of 2020, and run in early July 2020. The crawler is based on the open-source web browser Chromium and uses a custom instrumentation developed by using the Chrome debugging protocol (CDP) [11]. By connecting into its network tracing processes, we gather all the requests and responses performed by the browser during a web access. In order to avoid possible detections of our automated browser, we implemented the most recently-proposed methods [14, 51–53], also leveraged by other recent studies [47, 62]. When third-party scripts were loaded into each page we analyze the request, extract the destination domain and verify that the loaded entities were actually trackers by leveraging the tracker list used by Mozilla Firefox [37], and EasyPrivacy [15]. The two monitor different forms of tracking, such as web bugs, tracking scripts, and information collectors. Once the tracking domains are identified, we map the domain names to organizations based on three manually-curated lists: Disconnect [13], WhoTracks.me [12] and webxray [31].

We scanned the 2.33M websites in our telemetry using a server located in the US and discovered 6,320 distinct tracker names. To account for tracker variability due to geographic locations, we deployed additional crawlers in three different countries from three continents. For this, we leverage a commercial VPN service [6]. Specifically, we looked at browsing histories of users from France (6213), Brazil (5152), and Australia (5603), and crawled 130.70K, 67.81K, and 126.73K websites from the respective country. We report the results and compare them with the data collected from the US in Table 1. We found that on average 80.3% of the websites include exactly the same trackers, while another 6.9% has only one additional tracker. To obtain further insights into the remaining websites that have more than one different tracker ($\sim$12%), we compute the intersection over union (IoU) coefficient between the two sets of trackers obtained by crawling from US and the respective location: the rationale is that a result close to 1 (e.g., $>.8$) refers to very similar organization lists; on the other hand, a value close to 0 (e.g., $<.2$) implies the opposite. We finally assess that around 95.5% of the websites show no or subtle differences in the trackers detected, whereas we detect a diverse tracking ecosystem only on a very small subset of 4.5% domains. We dedicated appendix A to discuss the implications that the geographic location of the crawler has on our overall findings.

Table 1: Comparison summary between trackers detected crawling websites from US and France, Brazil and Australia

	US			
Country	Same trackers	$\pm$1 tracker	IoU >0.8	IoU <0.2
France	84.42%	5.52%	0.46%	4.97%
Brazil	79.28%	6.84%	1.14%	4.56%
Australia	77.20%	8.04%	1.84%	4.04%

3.3 Website categories and risk

By using the public classification service from the same security vendor detailed in Appendix C, we were able to assign a category to the websites in our telemetry. To better investigate the impact of tracking and the prevalence of different trackers on websites that could be related to user's sensitive information, we selected a set of *sensitive categories*: Health, Legal, Financial, Sexuality, Political, and Religion. Our decision was guided by categories defined as sensitive in various data protection laws [17, 18, 28], and used in recent studies [32, 48]. Finally, we additionally assigned a security-related risk level to each distinct website in the telemetry by leveraging the rating service from the security vendor described in Appendix D. For a given domain, the service outputs a score between 1 (completely safe) and 10 (certainly malicious).

3.4 Tracker relationships

A previous study [49] investigated the relationships among 810K actors during the creation and sharing of cookies

through cookie chains. In particular, the authors shed light on the role of those acting as dispatchers of information, receivers, or cookies direct creators.

We manually extracted the dependency relationships of the top trackers from the linkage graph and its related table in their manuscript, and used them to evaluate information sharing between a sender and a receiver organization. In this measurement, we assume that this happens in all the cases, i.e., the former always shares any data with the latter: although for many of the relationships this does not match the reality —trackers share part of the information and not for all the webpages—, in our discussion we consider it as an upper bound in order to evaluate the worst-case scenario for some of our findings.

4 Dataset Statistics

The users in our telemetry span 214 of the 249 countries with an assigned ISO 3166-1 code [2]. More than 44% of the users are located in North America (with 38% of them in the United States). Asia and Europe follow with about 20% of the users each. In South America, Africa and Oceania we find the lowest percentages (less than 17% overall). We report the complete geographical breakdown in Table 9 in Appendix B.

On average, the median user is active slightly less than 6 days out of 8, and for a number of hours per day that ranges from 3 to 10. We report a graphical summary of users' activity in terms of mean browsing days and hours in Appendix E.

We further look at the aggregated users' browsing behaviors in our dataset: we detect that on average during the 8 days, users present a history with 406 entries, browse 19 distinct categories, 118 different webpages, visit more than once 59 of them, and encounter 3,170 trackers from 177 distinct organizations. Additionally, we measure that 93% of them have less than 10 trackers, and for a single webpage visited, users encounter on average 3.5 different trackers.

In Table 2 we provide a summary of both sensitive and top-10 categories in our dataset, sorted by the number of websites they encompass. Webpages related to users' *Health* are the most frequent among the sensitive categories, also reporting the longest list of trackers encountered (34% of the 6,320 trackers). On the contrary, the *Political* category, the smallest among the sensitive category in terms of number of websites, visiting users, and different trackers detected, shows the highest average of trackers. This suggests that fewer organizations focus on political websites but more consistently. We will come back to this comparison in Section 6, when we will discuss in more detail which and how much sensitive information the different trackers can obtain about users. Regarding the other, non-sensitive, categories almost the totality of users browse websites classified in the *Technology/Internet* and *Business/Economy* groups: we indeed detect in the pages of these two categories almost 50% of the tracking organizations.

We finally analyze the coverage of the top 20 trackers in our dataset, reporting the percentage of known history, websites and users who encounter them in Table 3, together with the average values for all the trackers. We point out to the reader the subtle difference between two recurrent concepts throughout the manuscript: when computing the known history percentage by a tracker, we refer to the portion of entries in our telemetry in which we detect the tracker —thus also considering revisited websites across hours and days. On the contrary, when reporting the known website percentage, we only consider the fraction of unique website IDs —i.e., we do not take into account revisited webpages.

Table 2: Overview of sensitive (above) and top-10 (below) categories in our dataset

Category	% Websites	Avg Trackers	% Trackers	% Users
Health	4.89	10.80	34.78	33.89
Sexuality	2.89	2.82	24.75	17.97
Financial	2.00	7.77	29.11	53.86
Legal	1.95	2.64	19.73	34.62
Religion	1.91	8.29	20.41	19.84
Political	0.52	14.25	16.66	11.58
Business/Economy	11.62	8.64	48.94	83.30
Technology/Internet	6.55	9.36	46.06	99.18
Shopping	6.37	14.52	38.32	58.56
Education	4.44	7.01	30.97	50.68
Suspicious	3.79	1.45	28.84	40.49
Entertainment	3.47	13.84	40.41	53.34
Travel	2.76	8.27	31.33	33.36
Search Engines	2.43	3.43	26.41	94.32
Restaurants/Food	2.24	18.90	27.07	21.85
Personal Sites	2.18	8.90	26.61	19.66

At a glance, Google clearly stands out, being directly present in almost 73% of the websites in our dataset. The other top-20 tracking organizations cover on average 15.27% of users' history and 8.45% of the websites. From the users' perspective, almost all of them encounter at least once one of the top organizations in Table 3. Interestingly, while the average number of users reached by a single tracker is 3%, we measure that almost the totality encounters at least one tracker. The few exceptions – 419 users corresponding to 0.16% of the total – have a clean and not-tracked history. However, the small number together with the fact that those users only browsed an average of two different websites in 8 days, suggests that in practice *everyone* who browses the web is tracked to some extent.

It is also interesting to observe the difference between the two middle columns, i.e., the coverage in terms of unique websites and the one in terms of entries in the users' browsing history. Google is the only tracker in which the first is bigger than the second, meaning that it is the only company that also covers many less popular websites that do not receive many visits. Microsoft is instead an example of a company that seems to focus mostly on popular sites, as shown by the fact that its history coverage is more than five times the one of websites.

Table 3: Coverage overview for the top-20 companies involved in tracking in our dataset. The percentage of users' history (websites) without any trackers is 20.07 (23.11).

Tracker	% History	% Websites	% Users
Google	63.07	72.33	99.76
Facebook	30.05	26.53	98.33
Microsoft	22.97	4.11	97.56
Adobe	19.92	7.83	97.42
Appnexus	18.91	5.27	97.58
Yahoo!	17.36	5.33	97.05
Twitter	16.73	6.10	96.85
Rubiconproject	15.16	4.52	96.79
Thetradedesk	14.54	3.61	96.37
Rapleaf	13.92	4.19	96.12
Casalemedia	13.68	4.26	96.61
Pubmatic	13.30	4.08	96.45
Openx	13.09	4.16	96.40
Mediamath	12.69	2.49	96.56
Drawbridge	12.41	3.30	94.39
Amazon.com	12.00	2.69	95.14
Akamaitechnologies	11.53	1.11	95.48
Linkedin	11.33	2.09	94.13
Quantcast	10.84	3.68	95.81
Taboola	9.65	1.45	94.24
Average	0.14	0.06	3.00
Untracked	20.07	23.11	0.16

4.1 Dataset Limitations

Although our telemetry is large and contains hundreds of thousands of users from almost every region in the world, it may still be subject to some selection biases. For instance, it only includes users who protected themselves by installing an AV product and opted in to share their data: users who decided not to opt-in due to privacy concerns could behave differently, being more conscious with respect to tracking and high-risk websites. Furthermore, our entire telemetry comes from Windows machines. It is possible that users running other OSes (e.g., macOS and Linux) or browsing through mobile devices may exhibit a different behavior. Moreover, our data covers only 8 days of users' browsing experiences. As we will discuss in the following sections, users encounter the vast majority of the trackers already in the first day of browsing. Therefore, it is very unlikely that the final results would significantly get impacted with more data.

5 Standing in users' shoes

We start our analysis of web tracking by looking at the trends from the users' perspective. Our goal is to use our telemetry information to estimate how much, and how fast, real users encounter web trackers during their daily activity. We are also interested in finding whether some users are more exposed than others, or whether a certain class of online behavior leads to higher or lower privacy risks.

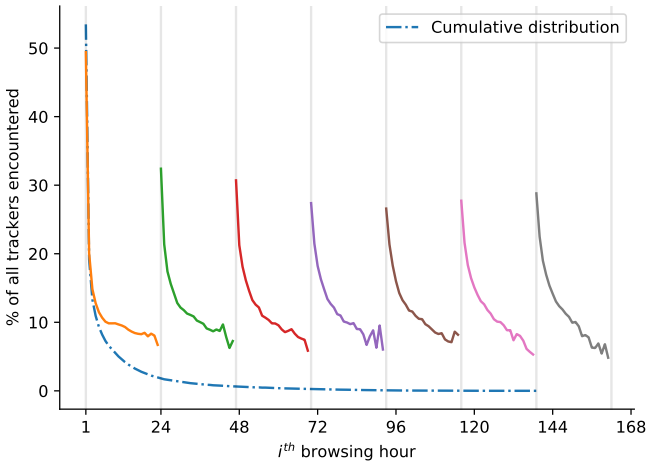


Figure 1: Cumulative and daily distribution of new trackers encountered per hour of activity. For the j^{th} daily curve, the i^{th} hours close to the boundary with the next day refer to users active i hours in the j^{th} day —thus active almost all the j^{th} day.

5.1 How long does it take for a user to encounter trackers?

To answer this first question we investigate the relationship between the time a user spends browsing the Web and the number of new trackers she encounters. To this end, we initialize a *cumulative* tracker set for each user. Then, for each cumulative i^{th} hour spent browsing, we add the new trackers encountered to the set and register its length variation from the previous time interval. Each i^{th} point of the blue curve in Figure 1 is then obtained by averaging the i^{th} values of all the users active at least i hours.

In a similar way, we maintain also a *daily* set for each user. For every j^{th} day, we add new trackers and register variations as for the cumulative case. We finally compute each of the i^{th} points for a j^{th} daily curve in Figure 1 by averaging the values of users active at least i hours in the j^{th} day. We do not include the daily plot of the 8th day in our telemetry because our data does not cover all its 24 hours.

The analysis of Figure 1 provides three important findings. First, the curve of new trackers per hour of activity follows a decreasing exponential distribution, with a drastic drop in the first 12 hours. Indeed, the average of new trackers encountered falls below 5 after 12 hours, below 2 after 22 hours and users encounter almost no new tracker after 35 hours of activity.

Another way to look at this data is to compute how many hours it takes for users to encounter a given percentage of all the trackers they encountered during the week under analysis (on average 177 trackers per user). In this case, on average after 2, 12, and 24 hours of activity users have already encountered respectively 50%, 84%, and 94% of their trackers.

The second interesting finding is that given a window of

i hours (e.g., 24), users who are active for more consecutive hours encounters a higher number of trackers with respect to the others. This discrepancy is clearly visible in Figure 1, when comparing the first part of the cumulative curve with the daily curve of the first day.

For instance, we can consider two users that both have three hours of activity over a 24h window. The first browses the Web in three separate sessions of one hour each – in the morning, afternoon, and evening. The second browses instead for three hours straight in a single session. In our experiments, we noticed that the second user is more likely to encounter a higher number of unique trackers. And the reason is that sessions that are far apart are more likely to have larger intersections in the visited websites. In other words, the likelihood of revisiting the same websites and running into already encountered trackers is higher in those cases. On the contrary, users characterized by longer browsing sessions show higher variability in the websites and trackers encountered.

The third observation we can make from Figure 1 is that all daily curves have really similar shapes, with a sudden decrease in the number of new encountered trackers in the very first hours. This suggests that, even if the user would restart with a clean browsing history every day, it would only take two hours on average to re-encounter 50% of all trackers. In other words, if a user encounters on average 177 different trackers per week, half of them are regularly encountered every day within the first two hours of web browsing.

So far we have captured the users' activity by counting the time they spend browsing. Another way to do that is to count the number of visited sites. The trend of how the newly encountered trackers evolves for each new website visited is summarized in Figure 2. The points on the blue curve are obtained by averaging the number of new trackers encountered for the i^{th} new visited website, among users who browse at least i distinct websites. The distribution in Figure 2 shows a similar trend of the corresponding cumulative curve when considering the hours of activity (Figure 1). The exponential shape has a maximum at 9 — suggesting that users encounter more than the average of 3.5 trackers when visiting the very first website, probably indicating a popular page with multiple trackers—, and quickly drops: after 20 different websites, users only encounter on average 2 new trackers. When computed in percentages, our data shows that by visiting 22, 100, and 300 distinct websites, the trackers encountered are respectively 50%, 75%, and 85% of the total encountered over the week.

However, this represents a *best-case scenario* that considers each tracker in isolation. In reality, trackers also exchange data with one another. Therefore, we complement our analysis by plotting a second curve, but this time considering the relationships among the different actors indicated in Section 3.4. In this case, when we add a new encountered tracker to the set, we also add all other trackers that directly receive information from it [49]. This curve, in orange in the graph, represents a *worst-case scenario*. In fact, the fact that

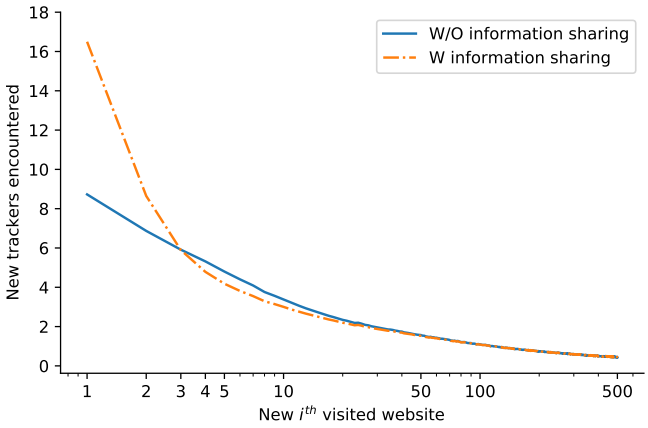


Figure 2: Average number of new trackers per new website

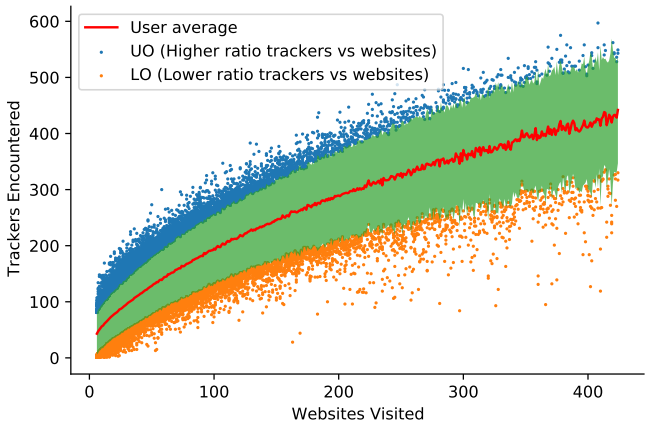


Figure 3: Correlation trend between the number of visited websites and encountered trackers

a relationship exists between two trackers does not imply that the two companies share *all* data about all users on all websites. Therefore, reality lies somewhere in between the two curves.

Even in the worst-case scenario, it is interesting to observe that the data shared among trackers exposes the users to a higher number of tracking companies for the first few visited websites. However, after around 20 websites the two curves overlap, showing that at that point the number of new trackers encountered by the user is independent from possible collaborations among trackers.

Summary: users find half of the trackers they are going to find during the full week just in the first hours and website visits (this pattern happens every day). Moreover, users who are active for more consecutive hours, tend to find a more variety of trackers.

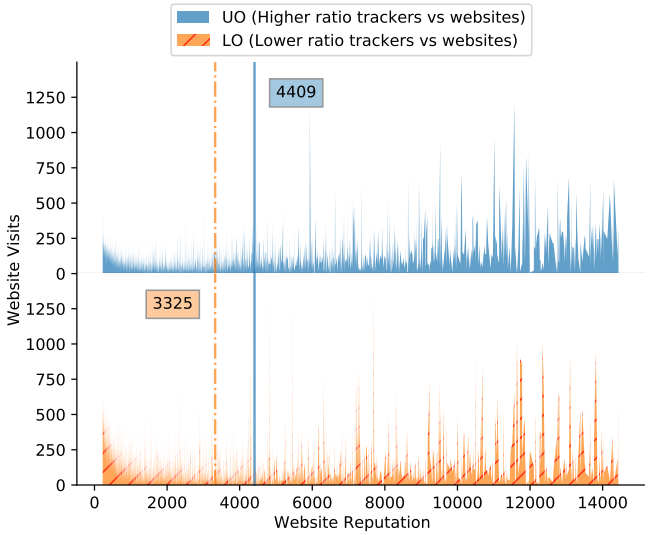


Figure 4: Website reputation distribution for *UO* and *LO*. The difference between the means in the two groups is significantly different $(Welch'st = 113.06, p < 0.001)$.

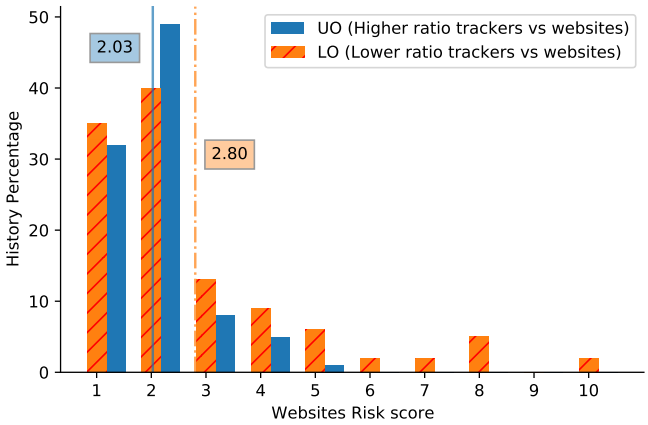


Figure 5: Website risk score distribution for *UO* and *LO* The difference between the means in the two groups is significantly different $(Welch'st = 432.41, p < 0.001)$.

5.2 Is there a correlation among distinct visited websites and encountered trackers?

We now look at the correlation between the total number of distinct websites visited by a user and the number of encountered trackers. In particular, we are interested in finding (and comparing) those users that encounter a disproportionate number of trackers despite visiting a few websites, and those that instead encounter a few trackers while visiting many different pages.

To begin with, we compute the two attributes (distinct websites and distinct trackers) and plot them for each user in Figure 3: a point (x,y) on the red curve represents the average number y of trackers encountered for users who visit x different websites, and the green area defines the 95% confidence interval.

The total number of visited websites positively correlates with the trackers encountered (Pearson Correlation Coefficient: 0.98, $p < 0.001$). However, Figure 3 exhibits two classes of outliers, whose attributes fall out outside the confidence interval boundaries. Specifically, we define *Upper Outliers (UO)* those with an abnormal-higher ratio between encountered trackers and visited websites (blue dots in the picture, users that encounter a lot of trackers while not visiting many websites). On the contrary, we report in orange the *Lower Outliers (LO)* (users that browse a lot but encounter less trackers), for which this ratio is lower than the average and outside the confidence interval. The UO and LO sets contain respectively 6,726 and 5,552 users, which together account for 4.6% of the users in our dataset.

To investigate whether any significant difference exists in the websites visited by the two groups of outliers, we use two metrics: popularity and security risk score. We compute the popularity of each website in our telemetry by simply considering the number of times it appears in different users' browsing histories. This score is conceptually similar to the reputation returned by online rating services [3–5, 7], and it is strictly related to the data in our experiment.

Given a popularity x, we separately plot for each group the sum of visits that each distinct website with reputation x receives (Figure 4). We next compute the weighted average for UO and LO according to the following criterion:

$$Wavg = \frac{\sum_{reput=1}^{max\_reput} reput * visits(reput)}{sum(visits)}$$

The two averages, represented by the vertical lines in the figure, show that users that encounter fewer trackers (LO group) are indeed visiting less popular websites. Instead, users who browse fewer websites but encounter on average more trackers mainly visit popular web pages: this is the case, for instance, of very popular news websites, social media, and online marketplaces, which incorporate a large number of advertisers, and a myriad of analytics services. For those users within the green zone in Figure 3, the reputation score falls between the one of UO and LO (i.e., 3,997), confirming our hypothesis that reputable sites are more tracked.

To compute the security risk score we leverage the website risk score provided by the AV vendor. Then, for each set of users, we split the websites they visited according to their risk value, and plot a histogram with the percentage of the total history they account for (Figure 5). The figure also includes the weighted average of both groups, computed by following the same procedure described for Figure 4. The plot shows that users in the UO group mainly browse benign websites. In our dataset, not a single website visited by these users had a rating that classifies it as either suspicious or malicious ($>=6$). On the other end of the spectrum, users in the LO group visit a larger percentage of dangerous sites. Similarly, the users in the green zone visit websites with low-risk scores however slightly higher than those UO users (2.6 risk score).

Table 4: Zero-Tracker website percentage and risk score for top and bottom 0-tracker categories

Category	% with zero trackers	Avg risk
Malicious Outbound Data/Botnets	90.23	9.40
Business/Economy	70.45	3.93
Potentially Unwanted Software	56.75	7.00
Spam	56.21	7.00
Placeholders	55.81	6.00
Suspicious	53.41	7.58
Scam/Questionable/Illegal	49.24	7.35
Email	43.69	4.47
Malicious Sources/Malnets	42.89	9.99
Social Networking	12.92	4.09
E-Card/Invitations	12.69	3.44
Informational	12.39	3.93
Alcohol	12.11	4.03
Translation	12.02	3.29
Restaurants/Food	11.85	4.19
Charitable Organizations	11.78	3.95
News/Media	10.93	3.77

Overall, we found that websites that include no trackers are often less popular and characterized by a higher security risk. Table 4 reports the top and bottom website categories, sorted by the percentage of webpages in which we do not detect any trackers. The top categories show a considerably higher risk score (6.96 on average) than the bottom (3.90 on average) suggesting that the former often present suspicious or malicious content rather than the latter (confirmed also by the category names). A clear exception in the top half of the table is represented by the *Business/Economy* category, which is both low-risk and low-tracking. This category represents websites devoted to businesses (including information and management) that are not linked to any selling activity. Taking this into account, a possible explanation is that websites in this group are directly related to customers or employees, so they do not include any type of tracking.

Summary: the more a user stays away from dubious websites, the more trackers she encounters. On the opposite, users that spend more time on less popular and high-risk sites are more exposed to security risks but less exposed to tracking.

5.3 How Frequently do Users Encounter the Same Trackers?

So far we only looked at how often users encounter new trackers. But the key point of tracking is identifying the same user across different websites. So, if a user encounters a specific tracker only once a day, then deleting its cookie at the end of the browsing session could prevent the tracker to connect the different visited sites. It is clearly possible that some trackers perform some type of browser fingerprinting [24] in order to be able to track users around. In these cases, deleting cookies would not avoid tracking. However, as cookies are still the de-facto tracking method on the web [49], we wanted to investigate how effective the cookie cleaning option could be to improve users' privacy posture.

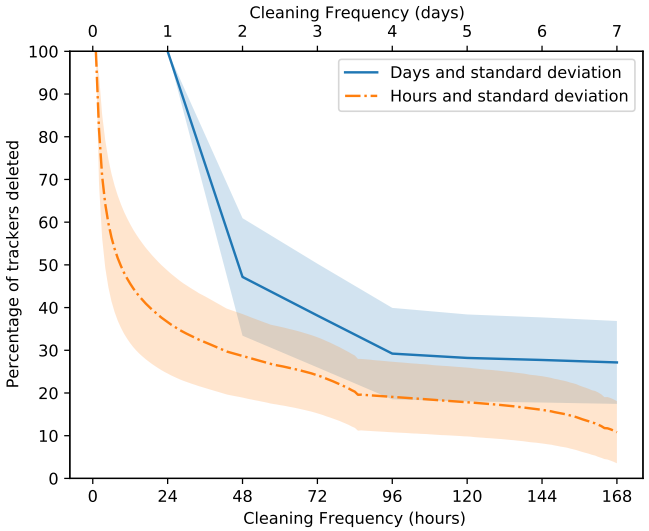


Figure 6: Percentage of trackers deleted according to the frequency (browsing hours and days) of cookie cleaning.

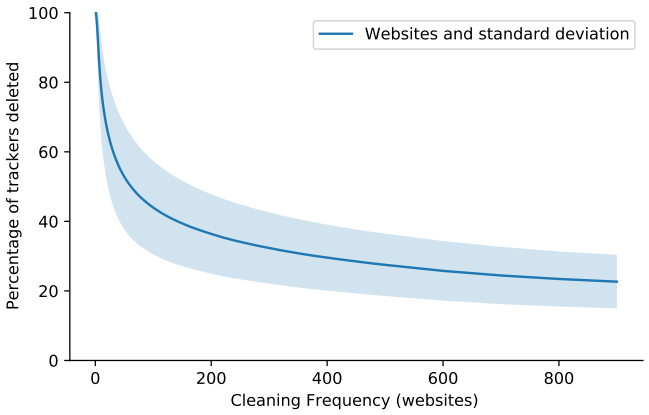


Figure 7: Percentage of trackers deleted according to the frequency (browsed websites) of cookie cleaning.

To better understand this aspect we looked at how frequently each tracker was encountered by each user. In Table 5, we report the percentage of users for which the top-5 most recurrent trackers appear with a frequency lower than 2 hours. Google, for instance, is encountered on average every 1.11 hours. This means that to fully prevent the largest company in our dataset from being involved in tracking practices, a user should delete the cookies after every single browsing hour, which is obviously not realistic. Figure 6 and Figure 7 respectively report the cumulative distributions for a time-based and site-based perspective. The plots show that 50% of the trackers are

Table 5: Top-5 trackers according to the frequency (browsing hours) of appearance. % Users refers to users for which the tracker appears with a frequency < 2 browsing hours.

Tracker	% Users	Avg frequency (hours)
Google	80.41	1.11
Microsoft	67.61	1.29
Twitter	67.18	1.41
Yahoo!	66.25	1.43
Rubiconproject	62.68	1.44

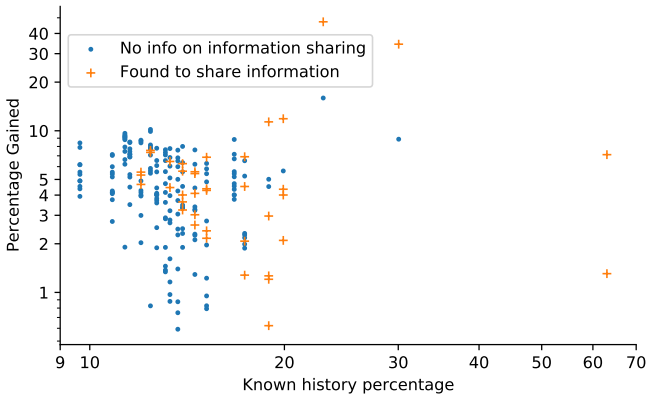


Figure 8: Possible browsing history gain through collaboration

repeatedly encountered every 8 hours or 60 websites. In other words, if the cookies are cleaned up every 8 hours or after 60 website visits, only half of the trackers could be prevented from tracking. However, cookie cleaning is clearly not an absolute solutions for those privacy conscious users who do not want to be tracked by any means: this practice is not effective against big players that can know much more and are encountered much frequently (on average every 1.34 browsing hours).

Summary: Users encounter on average 177 trackers during one week. In addition, 18.31% of the users encounter trackers every hour and 1.73% encounter trackers every website.

6 The Knowledge of Trackers

In the previous section, we have seen that the average users encounter 84% of the trackers within just half a day of web browsing. While this is very concerning for the privacy of Internet users, the impact on their privacy might not be as significant and worrying unless those trackers can compromise a significant fraction of the users' browsing history. In this section, we take a closer look to estimate how much information about users is known (or potentially known) by those trackers. We first assess to what extent main trackers on the visited websites know about the users' browsing histories, and then how much additional coverage they could gain by sharing information among one another. We also investigate the type of information that could be learned about the identity of users through regularly browsing particular types of websites. Finally, we conclude the section with an optimal tracking strategy analysis.

6.1 How much do trackers know about you?

For each tracker we identified in our dataset, we computed the average fraction of browsing history known, the percentage of websites in which they are present, and also the fraction of users who encounter them. On average, each tracker tracks 3% of the users and knows 0.14% of their browsing history. However, the top trackers (such as Google, Facebook, and Microsoft) are quite far from the average. In fact, they are able to track nearly all users, as can be seen from Table 3, and they know on average 47% of each user's browsing history. Google alone, which is the biggest player in the tracking ecosystem, covers 64% of the average users' history logs. The percentage increases to 80% for 9.73% of the users, and reaches a stunning 100% for 2% of them.

Summary: Large trackers know, on average, nearly half of the browsing history of almost all users. For roughly 10% of the users in our dataset Google alone was tracking over 80% of the visited websites.

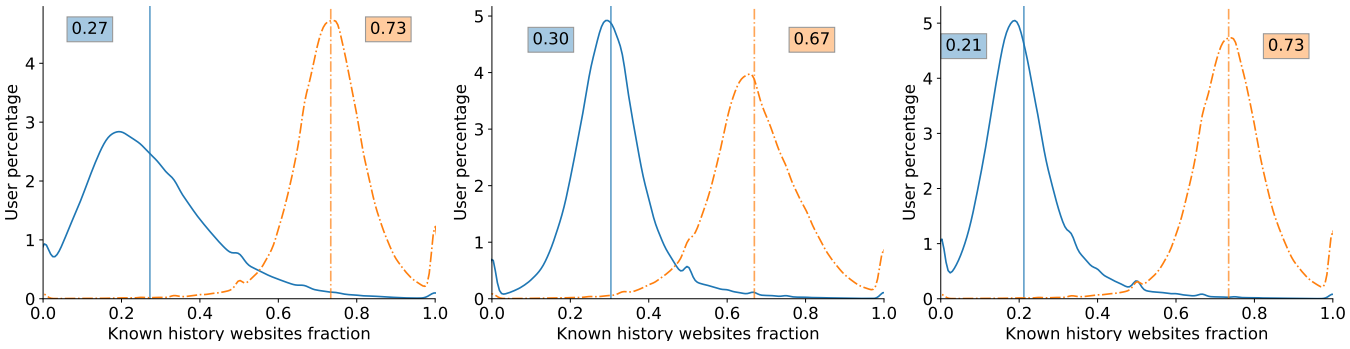
6.2 How much can trackers know about you through collaboration?

Collaboration among trackers is not a new phenomenon [19,41, 49]. It allows them to merge the user data with another tracker, reconstructing users' browsing history, and bypassing the same-origin policy [60]. In order to do it, tracking companies can use multiple methods, with cookie sharing/synchronization being the most common one. For example, a tracker can include its cookie in the request of another third party, facilitating an information-sharing channel even if not directly present in that specific website. Our goal here is to estimate the concrete impact of such collaborations on users' browsing history, which was not explored before by other studies.

In the previous section, we have seen that with the exception of Google, none of the other trackers knows more than 30 percent of the average user's browsing history. Clearly, if Google shared its knowledge with any other tracker, they could also achieve similar coverage. However, this is not a very realistic scenario from a strategic point of view. On the other hand, collaboration among smaller players in the ecosystem might make more sense. Therefore, to understand how much information trackers could gain through collaboration, we calculated the browsing history gain for all possible pairs of companies among the top 20 trackers in our dataset and plotted the percentage of gain versus known history percentage in Figure 8. If the two companies were already known to collaborate according to previous measurements [49], we colored them in orange. If you remove the top three players, in general most trackers over the top 20 can know between 10 and 20% of the browsing history

Table 6: Upper (Lower) part: top and bottom 5 relationships sorted by ascending overlapping (descending gain)

Tracker A (Receiver)	Tracker B (Sender)	Coverage Tracker A	Coverage Tracker A+B	Gain	% B overlapping
Linkedin	Amazon.com	11.33	20.62	9.29	22.57
Amazon.com	Linkedin	12.00	20.62	8.62	23.90
Microsoft	Google	22.97	70.19	47.22	25.13
Taboola	Linkedin	9.65	17.54	7.89	30.40
Linkedin	Openx	11.33	20.43	9.10	30.49
		...			
Rubiconproject	Casalemedia	15.16	15.95	0.79	94.21
Casalemedia	Openx	13.68	14.43	0.75	94.28
Casalemedia	Pubmatic	13.68	14.27	0.59	95.55
Google	Facebook	63.07	64.38	1.31	95.65
Appnexus	Rubiconproject	18.91	19.54	0.62	95.89
Microsoft	Google	22.97	70.19	47.22	25.13
Facebook	Google	30.05	64.38	34.33	45.56
Microsoft	Facebook	22.97	38.92	15.95	46.91
Adobe	Microsoft	19.92	31.79	11.87	48.30
Appnexus	Microsoft	18.91	30.27	11.35	50.56
		...			
Drawbridge	Linkedin	12.41	13.24	0.83	92.71
Rubiconproject	Casalemedia	15.16	15.95	0.79	94.21
Casalemedia	Openx	13.68	14.43	0.75	94.28
Appnexus	Rubiconproject	18.91	19.54	0.62	95.89
Casalemedia	Pubmatic	13.68	14.27	0.59	95.55



(a) Microsoft (b) Facebook (c) Appnexus

Figure 9: Known-history percentage distribution of the trackers that directly appear the most in users' history without (solid line) and with information sharing (dotted line). The percentage of users' history without any tracker is 20.07%.

of the users. Through collaboration, they can increase their knowledge of an additional 5 to 10% (mean gain is 5.3%) in the best case scenario unless they can collaborate with Google.

In Table 6 we also provide concrete examples for some of the interesting collaboration options. Similarly, those collaborations that are known to exist by other means are marked in gray. The most obvious gain examples come from the collaboration among the biggest players. Because in most of the websites in which we observe Facebook, we also encounter Google (95.65%), Google gains not much (1.31%) from getting information from Facebook. However, Facebook could immensely increase its knowledge, up to 64.38%, from a potential collaboration with Google. Another interesting observation is that Microsoft and Google do not target similar sets of websites, therefore a possible collaboration would have a much larger impact. On the contrary, the overlap among the top 20 trackers ranges between 23 and 96% (mean overlap of 64%). This clearly indicates that many of them are tracking users in a very similar set of websites.

Now let's look at the worst-case scenario, in which we assume that all trackers that were identified to be sharing information according to recent studies (see Section 3.4) collaborate to increase their knowledge as much as possible. In Figure 9, we provide three examples of how much information

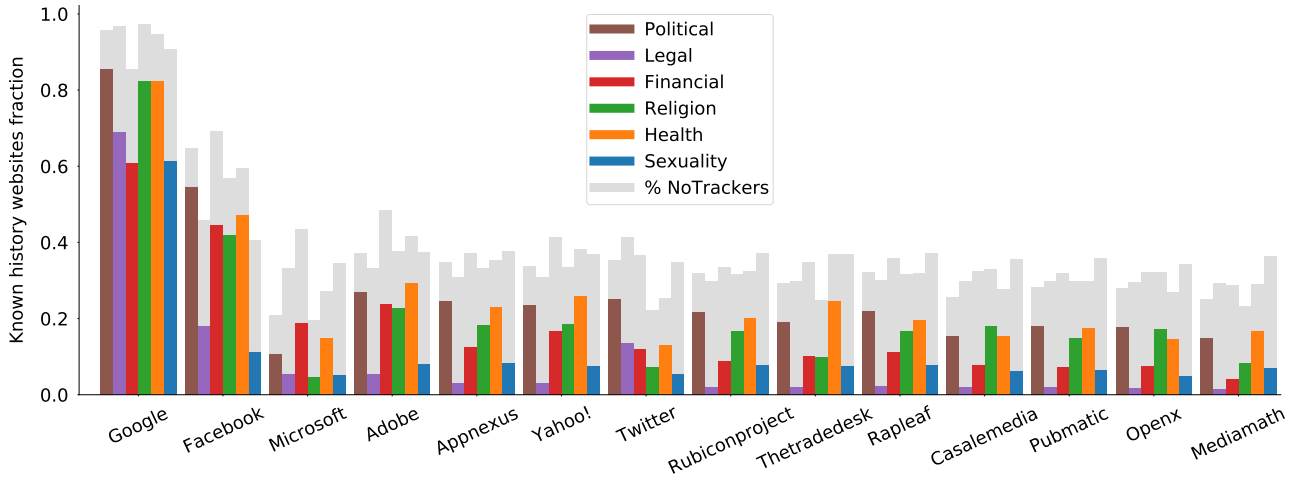


Figure 10: Known history percentages of the 6 sensitive categories by the top trackers.

can be potentially gained in such a scenario. It is interesting to see that Microsoft could potentially already know up to 73% of the users' browsing history (instead of the 27% it has if it was completely disconnected from other players). Another similar spike is observed in Appnexus (from 21 to 73%). While the gain for Microsoft is mostly due to its relationship with Google, Appnexus receives information from a variety of other trackers including Microsoft, Adobe, Yahoo!, and more. Again, these numbers assume a complete share of all tracking information among the companies, so in reality the numbers are likely somewhere in between the two scenarios (no collaboration and full collaboration).

Summary: Top trackers overlap on 64% of the websites where they track users. However, they can gain an additional 5% and 10% in their history coverage through collaborations. The gain increases to up to 50% if multiple trackers share data.

6.3 What type of sensitive information can be obtained about you?

Visiting or regularly browsing particular types of websites could reveal sensitive information about users. In this part of our analysis, we focus on websites that could fall into sensitive categories and check which trackers are present on those sites and could therefore gain access to private users' information. In particular, we identified six categories that are widely considered to be sensitive (see Section 3.3) and we computed the portion known by top trackers. Figure 10 reports the averages over the whole dataset. In gray, we represent the percentage of history in which we do not detect any trackers.

At a glance, we observe that the tracking activity is not uniform among the six sensitive categories: while the percentage of untracked history is very low in the *Health*, *Religion*, and *Political* categories (respectively 12, 15 and 10%), the fraction doubles for the *Sexuality*, *Financial*, and *Legal* classes (30, 24, 28%).

A first interesting case is the *Political* category: although it presents the lowest number of websites and users who browse it (see Table 2), it turns out to be the category the top trackers know the most about. In fact, our crawler detects multiple trackers on average on each of these pages, with top trackers uniformly present on most of them.

The *Legal* category results in the opposite case: top organizations on average know less than 5% of sites in this category, with the exclusion of Google (69.10%): we measure an average presence of 2.64 trackers for websites in this group.

More concretely, if looking at the per-tracker details in the graph, the figure presents similar trends and known history percentages, except for Google and Facebook. Since in general Google knows over 60% of the users' history, it is not very surprising that it also covers a good fraction of the browsing history related to the sensitive categories. For example, the Facebook case is utterly interesting. On the general data, it only knows up to 30% of the users' browsing history, which is in line with other top players. Despite that, it covers almost 60% of the browsing on the *Political* sector, and around 50% of the *Health* category. This seems to indicate that Facebook puts a particular effort in tracking specific website classes. On the other side of the spectrum lies Microsoft, which on the general data has a much larger coverage (over 20%) than its presence on sensitive website categories.

We also investigate whether the prevalence and tracking of sensitive websites are uniform across continents. For each of them, in Table 7 we report the average percentage of browsed websites per sensitive category together with the average number of trackers encountered. Results show no substantial differences across continents and confirm that sensitive information about *Health*, *Religion*, and *Political* is more subject to tracking practices, although their prevalence is very small in users' histories. The only comforting difference is observed in Europe. Very likely thanks to the GDPR, the average number

Table 7: Sensitive website prevalence and their average number of trackers in users' history. Values higher than the mean for all the categories are underlined.

Continent	Websites Percentage							Average trackers						
	All	Sexuality	Health	Religion	Financial	Legal	Political	All	Sexuality	Health	Religion	Financial	Legal	Political
Africa	**6.03**	1.74	0.84	0.49	1.45	0.70	0.11	**7.01**	4.50	10.34	7.50	6.64	2.53	7.53
Asia	**5.84**	0.96	0.89	0.42	2.07	1.27	0.09	**6.96**	5.96	14.44	7.31	5.54	1.79	7.39
Europe	**5.45**	1.91	0.87	0.50	1.80	0.86	0.12	**7.12**	5.56	7.52	5.87	4.81	2.82	6.10
North A.	**5.00**	1.27	1.20	0.75	3.12	0.77	0.18	**8.55**	6.79	12.97	8.68	8.23	3.70	14.69
Oceania	**4.85**	1.72	1.00	0.52	2.36	0.88	0.11	**7.69**	6.59	12.47	8.84	6.33	2.60	8.52
South A.	**5.38**	1.04	1.26	0.42	1.95	1.93	0.14	**7.20**	6.02	10.06	9.44	4.87	2.23	12.85

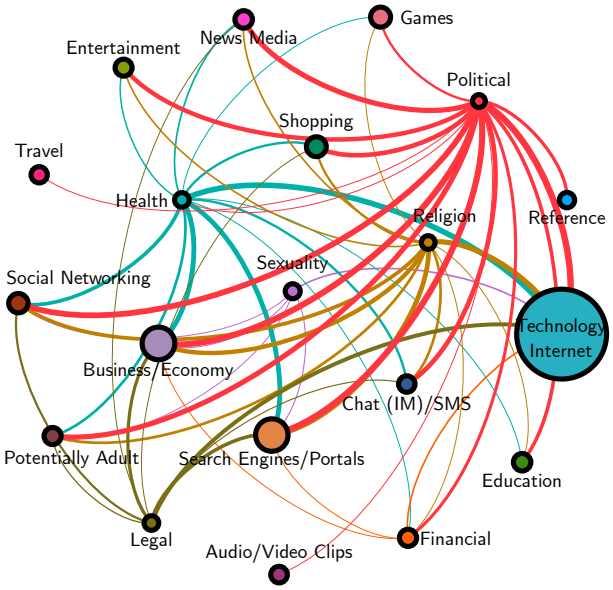


Figure 11: Relationships among sensitive and top categories

of trackers found in websites is lower than others.

As a next step, we investigate how much more information can be identified about a user's identity by connecting the pieces. For example, if a tracker knows that a user follows a particular political party or religious belief, can we estimate the likelihood of them knowing also about the user's travel plans, health interests, etc? To this end, we build a linkage graph among the sensitive categories and other website categories. We consider each user at the time, and isolate the history containing webpages of the sensitive category (*SI*) under analysis from the remaining part (*RI*) — note that the group also contains other sensitive categories besides the one we investigated so far. For each webpage in *SI*, we extract the list of trackers and check their presence in the remaining webpages of *RI*. Given the list of matched websites, we detect their categories and increase a counter for each of them. Once n webpages in *SI* have been analyzed, we divide each of the category counters by n, obtaining a ratio. For a single user, a ratio close to 1 between a sensitive category a and another one b means that, each time we encounter a website in a, the trackers also know that the user visited b.

We plot the resulting linkage graph in Figure 11. Node sizes represent the percentage of history that falls in the category: the biggest category is *Technology/Internet* (39% users' history), the smallest is *Political*, accounting for 0.13%. Each edge between two nodes expresses the average category correlation for all the users in our dataset. To increase the readability, the graph only includes the sensitive and the most prominent ten categories that have at least one ingoing edge with a weight greater than 70%. We observe that the strongest correlation percentage (95.55%) holds between *Political* and *Technology/Internet*, while the weakest (70.01%) between *Legal* and *Chat (IM)/SMS*.

We also see that some categories are much less connected with the others. For instance, *Sexuality* and *Financial* have very few connections with other categories, and those connections are very small. On the other hand, *Political* has many strong connections with many other categories found in the dataset. In the middle, we find cases like *Health*, *Religion* and *Legal*, that despite having more connections than the first two, only have a couple of strong connections with others. We also verify how the linkage graph varies according to users' geographical location, and find that relationships are stable across continents except from Asia, in which we see *Health* has stronger connections than *Political*.

Another interesting point is that sensitive categories do not seem to have many connections among them. However, we have to note that not having a direct connection in the graph does not necessarily indicate that trackers could not connect them through their relations to other common categories. For example, both *Political* and *Health* are connected to *Potentially Adult*, which could be used as a hub.

Summary: The trackers coverage for sensitive categories ranges between 10 and 30%. Even if these categories are not connected much between one another, some are strongly connected to other general categories such as *Potentially Adult* or *Entertainment*. Some trackers seem to focus on some particular sensitive categories and have 30% more coverage in those categories than on other websites.

Table 8: Top-10 prevalence in the 5K *key-websites*: trackers (left) and categories (right)

Tracker	% key websites
Google	66.04
Facebook	35.50
Adobe	21.54
Appnexus	19.02
Yahoo!	18.44
Microsoft	17.04
Rapleaf	16.00
Thetradedesk	15.56
Drawbridge	15.50
Rubiconproject	14.90

Category	% key websites
Technology/Internet	22.96
Business Economy	12.94
Shopping	6.60
News Media	5.82
Travel	5.04
Entertainment	3.90
Games	3.20
Suspicious	3.04
Financial Services	3.02
Education	2.70

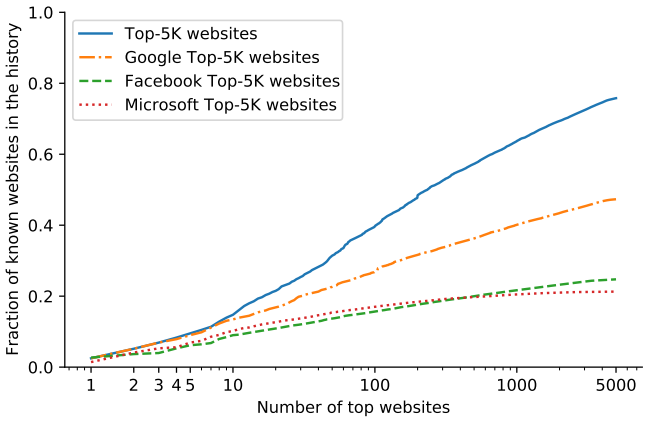


Figure 12: Optimal tracking strategy on *key-websites* vs top-3-tracker strategy on their top-5K websites

6.4 What is the optimal tracking strategy?

Earlier in this section, we have made estimations on how much browsing history knowledge could be obtained through collaboration among trackers, concluding that unless collaboration happens with Google, it is hard to gain a significant fraction of the browsing histories. An alternative option for the trackers to achieve the same goal is to plant themselves on *key websites*. For an optimal tracking strategy, the trackers need to build a list of *popular websites* such that the minimum number of them is required in order to cover a certain percentage of the whole users' history. To assess the effectiveness of this option, we created a sorted list of the 5K most reputable websites, according to the definition provided in Section 5.

In Figure 12, we plot how the known history percentage grows in relation to how many key websites the trackers need to work with. We also plot the existing presence of the top three trackers on those top 5K sites. The blue curve shows that, by cherry-picking only 200 websites, a company could observe 50% of the users browsing history. This value increases to 65% and 78% when extending the set of key webpages to the first 1K and 5K respectively (over a total of 2.33M in our dataset), indicating that being able to add a tracker to the top sites brings much more additional information than collaborating with other trackers.

As seen in Table 8, top players already show a significant presence on the *key websites*. When we look closer at the tracking strategy of Google, Facebook, and Microsoft (Figure 12), we identify interesting differences. First, although present in 66% of the key websites, the coverage of Google diverges from the optimal curve after considering only 10 websites (0.2% of the 5K): a sign that its presence is more prominent in the less reputable website of the group. We also noticed that Microsoft had a better coverage strategy than Facebook: although the two organizations show a similar trend in Figure 12, the first is only present in 17% of the key websites —half the percentage of the second—, suggesting that it appears in more reputable websites. In Table 8, we report the breakdown of categories together with the fraction of key websites for each of them: *Technology/Internet* and *Business/Economy* group a sheer number of webpages, being the two most popular categories overall.

Summary: Trackers could know 50% of the browsing history of the users in the dataset if they were present in just 200 websites. This number can increase to 78% with only 5K pages (out of 2.33M in our dataset). However, even the top trackers do not seem to follow this ideal strategy.

7 Comparison and Discussion

While web tracking is widely considered a common phenomenon, the results that we obtained by studying web tracking from the users' perspective show that it is considerably more widespread than previously thought. Previous studies [16, 25, 49, 50] attempted to quantify its scale by conducting large-scale measurements on open datasets, such as Alexa 1M [3] or Tranco [44]. However, while one would expect that crawling the most popular websites should provide an upper bound approximation of exposure, we found this to be wrong. For example, Google was found to track user activities on 46% [25, 50] of the top domains, but our study reveals that the actual knowledge of the users' histories reaches 73%. In the same way, Facebook prevalence was estimated around 18% [49], but our measurement shows it to be almost twice that value.

One of the main results of our study is to show that if the impact of web-tracking is measured only by considering top websites, the fraction of known browsing history would be largely under-estimated. Moreover, the relationship among the two is not always the same. As an example, Microsoft and Pubmatic appear both in 4% of the analyzed websites, but the former covers on average almost twice the users' browsing history compared to the latter (Table 3). The use of telemetry makes it also possible to quantify the exact impact of collaborations among organizations on end users. Previous studies discovered that 66% of the top-100 trackers share cookies [16] and that users with a larger browsing profile are tracked by

more identifier sharing domains [19]. Thanks to our analysis we now know that this practice could increase the knowledge that trackers have of the users' activity by almost 50%.

Another advantage of our method with respect to previous works is that it also allows us to shed light on the timing and frequency with which users are tracked, thus unveiling insights on research areas that have never been explored so far and whose investigation is impossible by crawling top websites. For instance, we show that users encounter almost all the tracking organizations in just half a day of activity. Even more worryingly, we show that the frequency with which some of the top trackers are encountered makes it infeasible to prevent their monitoring by simply deleting the cookie history.

The knowledge that tracker organizations have of users' browsing interests, habits, recurrence, location and hourly activity enables the creation of powerful profiles that get more and more refined and available to many players willing to purchase them. As a result, users risk to lose control of their private information and face several serious consequences. For example, a known use of tracking is the personalization of search results based on users' interests and the creation of the so-called Filter Bubble [65], a personalized search where an algorithm guesses what results the user would like to see based on previously collected information. Web tracking is also massively used to serve targeted advertising, facilitate marketing, and increase sales profit by influencing customer purchasing behaviors. In this respect, tracking can be used to modify product prices according to the geographical location and the financial situation of potential customers [9, 63]. Many companies also leverage this information to assess users' financial credibility [58, 59] and establish insurance coverage [22].

7.1 What can users do to protect themselves?

As web-tracking closely concerns users and their activities, several tools and strategies exist to defend against this practice, being the most important: cookie clearing, list/rules-based blocking, and network-level masking.

Cookie clearing – In order to significantly reduce cookie-based tracking, users could delete the cookies stored in their browsers. However, this approach is complex to strictly follow in the long term and it would require a lot of effort: users must delete cookies with high frequency (i.e., less than one hour according to our findings in Section 5) and cherry-pick the ones to delete in each case.

List/rules-based blocking – The most common solution is the use of browser extensions or privacy-centered browsers that maintain an up-to-date list of tracking domains or rules and block all the connections towards them, thus preventing data collection about browsing sessions. Some of them rely on large-scale crawls to analyze how the ecosystem evolves [13], and some others principally have a crowdsourcing model [15]. These kind of solutions are easy to setup (i.e., install and forget) and avoid the need for manually deleting cookies on a regular basis. However, blocking resources can sometimes generate unexpected functionality problems in the page. In order to avoid them, solutions generally offer a page-specific disable option, but as indicated in Section 6, a large percentage of the browsing history in sensible categories is being tracked by multiple companies, so users should be extremely careful when disabling protection tools in them.

Although these solutions exist, and are practical and effective, extension or application-based blocking is not yet widely adopted: privacy-centered browsers only represent 7.74% of the market, and only 8.5% of the users adopt tracker-blocking tools [26, 64]. Therefore raising awareness about the extent of web-tracking is crucial to increase these percentages and we believe that the quantitative insights presented in this work could be immensely helpful to serve this purpose.

Network-level masking – Section 4 shows that the knowledge of tracker organizations spans a high percentage of the users' browsing history, reaching up to 63% in the case of Google. Therefore, some protections can be implemented at the network level to protect a larger portion of users and devices. Protective measures can be installed in home routers [43] or adopted as a privacy layer in companies. Despite being flexible and allowing protection of multiple devices at the same time, those tools are more difficult to set up, and require users to regularly maintain them, discouraging regular web users in adopting them.

There are also some solutions to mask the user's real IP address from the remote site, thus preventing IP-based tracking. This goal is achieved by using anonymous proxy servers (which act as intermediary and offer anonymization services by removing sensitive information), virtual private networks (VPNs) (whose nodes result to be hosts of a single network, regardless of their physical locations), or Tor [1] (whose browser prevents tracking by routing the traffic through a chain of relays which protects the real user's IP address). Even if this is only one part of online tracking, some studies have already proven that a large percentage of users retain their same IP addresses for more than a month [35], allowing companies to use it as an identifier. When adopting this type of solutions, users should additionally use a list/rules-based blocking tool on top, to also avoid general types of tracking.

8 Conclusions

Despite the existence of these solutions and the users' awareness of online tracking practices, the adoption of such countermeasures is still limited. A possible reason is that users might feel they are not directly impacted. Our goal is to provide a more accurate measure of how web-tracking directly impacts them, and with evidence about how their online privacy is affected. We hope our findings can enable better decision making and foster a larger adoption of existing privacy-preserving services.

Acknowledgment

We would like to thank our shepherd Deepak Kumar for his help in significantly improving this paper, and all the anonymous reviewers for their constructive feedback. This project was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101019206 (Testable).

References

[1] TOR. https://www.torproject.org/download/.

[2] ISO 3166-1. https://en.wikipedia.org/wiki/ISO_3166-1, 1997. Accessed: 2021-02-04.

[3] Alexa, "seo and competitive analysis software.". https://www.alexa.com, 2021.

[4] Cisco umbrella, "umbrella popularity list.". https://umbrella-static.s3-us-west-1.amazonaws.com/index.html, 2021.

[5] Majestic, "the majestic million.". https://majestic.com/reports/majestic-million, 2021.

[6] Mullvad. https://mullvad.net, 2021.

[7] Quantcast, "audience insights that help you tell better stories.". https://www.quantcast.com/top-sites, 2021.

[8] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, and Bart Preneel. FPDetective: dusting the web for fingerprinters. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2013.

[9] Paul Belleflamme and Wouter Vergote. Monopoly price discrimination and privacy: The hidden cost of hiding. *Economics Letters*, 149:141–144, 2016.

[10] Sarah Bird, Ilana Segall, and Martin Lopatka. Replication: Why we still can't browse in peace: On the uniqueness and reidentifiability of web browsing histories. In *Symposium on Usable Privacy and Security (SOUPS)*, 2020.

[11] ChromeDevTools. DevTools Protocol API. https://github.com/ChromeDevTools/debugger-protocol-viewer, 2020.

[12] Cliqz GmbH. WhoTracks.me: Bringing Transparency to Online Tracking. https://github.com/cliqz-oss/whotracks.me, 2019.

[13] Disconnect. Make the web faster, more private, and more secure. https://github.com/disconnectme, 2019.

[14] Dymo. Missing Accept_languages in Request for Headless Mode. https://bugs.chromium.org/p/chromium/issues/detail?id=775911, 2017.

[15] EasyPrivacy. Easyprivacy filter subscription. https://github.com/easylist/easylist/tree/master/easyprivacy, 2020.

[16] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016.

[17] Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009. *Official Journal of the European Union*, 2009.

[18] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, 2016.

[19] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Tracking personal identifiers across the web. In *Conference on Passive and Active Network Measurement (PAM)*, 2016.

[20] Xuehui Hu, Guillermo Suarez de Tangil, and Nishanth Sastry. Multi-country study of third party trackers from real browser histories. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 70–86. IEEE, 2020.

[21] IAB. The socioeconomic impact of internet tracking. https://www.iab.com/wp-content/uploads/2020/02/The-Socio-Economic-Impact-of-Internet-Tracking.pdf, 2020.

[22] Privacy International. Social media intelligence and profiling in the insurance industry.... https://medium.com/privacy-international/social-media-intelligence-and-profiling-in-the-insurance-industry-4958fd11f86f, 2017.

[23] Internet Archive. Wayback machine. https://archive.org/, 2020.

[24] Umar Iqbal, Steven Englehardt, and Zubair Shafiq. Fingerprinting the fingerprinters: Learning to detect browser fingerprinting behaviors. *arXiv preprint arXiv:2008.04480*, 2020.

[25] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M Pujol. Whotracks. me: Shedding light on the opaque world of online tracking. *arXiv preprint arXiv:1804.08959*, 2018.

[26] Kinsta. Global desktop browser market share for 2020. https://kinsta.com/browser-market-share/, 2020.

[27] Balachander Krishnamurthy and Craig Wills. Privacy diffusion on the web: a longitudinal perspective. In *The Web Conference (WWW)*, 2009.

[28] Issie Lapowsky. California unanimously passes historic privacy bill. Wired, 06 2018.

[29] Yan Lau. A brief primer on the economics of targeted advertising. Technical report, Technical report, 2020.

[30] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *USENIX Security Symposium*, 2016.

[31] Tim Libert. Webxray, a tool for analyzing third-party content on webpages and identifying the companies which collect user data. https://github.com/timlib/webXray, 2019.

[32] Srdjan Matic, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. Identifying sensitive urls at web-scale. In *Proceedings of the ACM Internet Measurement Conference*, pages 619–633, 2020.

[33] Jonathan R Mayer and John C Mitchell. Third-party web tracking: Policy and technology. In *IEEE Symposium on Security and Privacy (Oakland)*, 2012.

[34] William Melicher, Mahmood Sharif, Joshua Tan, Lujo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. (do not) track me sometimes: Users' contextual preferences for web tracking. *Proceedings on Privacy Enhancing Technologies*, 2016(2):135–154, 2016.

[35] Vikas Mishra, Pierre Laperdrix, Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Martin Lopatka. Don't count me out: On the relevance of ip address in the tracking ecosystem. In *The World Wide Web Conference (WWW)*, 2020.

[36] Rani Molla. Advertisers will spend $40 billion more on internet ads than on tv ads this year. Recide, 2018.

[37] Mozilla Foundation. Security/Tracking protection. https://wiki.mozilla.org/Security/Tracking_protection, 2020.

[38] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proceedings of IEEE Symposium on Security and Privacy (Oakland)*, 2013.

[39] NortonLifeLock. Nortonlifelock global privacy statement. https://www.nortonlifelock.com/us/en/privacy/global-privacy-statement/, 2021.

[40] Lukasz Olejnik, Claude Castelluccia, and Artur Janc. Why johnny can't browse in peace: On the uniqueness of web browsing history patterns. 2012.

[41] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference (WWW)*, 2019.

[42] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P Markatos. The cost of digital advertisement: Comparing user and advertiser views. In *The World Wide Web Conference (WWW)*, 2018.

[43] Pi-Hole. Network-wide ad blocking. https://pi-hole.net/, 2021.

[44] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156*, 2018.

[45] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *Networked Systems Design and Implementation (NSDI)*, 2012.

[46] Sonam Samat, Alessandro Acquisti, and Linda Babcock. Raise the curtains: The effect of awareness about targeting on consumer attitudes and purchase intentions. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[47] Iskander Sanchez-Rola, Davide Balzarotti, Christopher Kruegel, Giovanni Vigna, and Igor Santos. Dirty Clicks: a Study of the Usability and Security Implications of Click-related Behaviors on the Web. In *The World Wid Web Conference (WWW)*, 2020.

[48] Iskander Sanchez-Rola, Davide Balzarotti, and Igor Santos. BakingTimer: Privacy Analysis of Server-Side Request Processing Time. In *Annual Computer Security Applications Conference (ACSAC)*, 2019.

[49] Iskander Sanchez-Rola, Matteo Dell'Amico, Davide Balzarotti, Pierre-Antoine Vervier, and Leyla Bilge. Journey to the Center of the Cookie Ecosystem: Unraveling Actors' Roles and Relationships. In *Proceedings of IEEE Symposium on Security and Privacy (Oakland)*, 2021.

[50] Iskander Sanchez-Rola and Igor Santos. Knockin' on trackers' door: Large-scale automatic analysis of web tracking. In *Proceedings of the International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*, 2018.

[51] Evan Sangaline. Making Chrome Headless Undetectable. https://intoli.com/blog/making-chrome-headless-undetectable/, 2017.

[52] Evan Sangaline. It is Not Possible to Detect and Block Chrome Headless. https://intoli.com/blog/not-possible-to-block-chrome-headless/, 2018.

[53] Evan Sangaline. Bypassing Headless Chrome Tests, the game goes on... https://www.tenantbase.com/tech/blog/cat-and-mouse/, 2019.

[54] Sarah Berry. 2020 search market share: 5 hard truths about today's market. https://www.webfx.com/blog/seo/2019-search-market-share/, 2020.

[55] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. Clash of the trackers: measuring the evolution of the online tracking ecosystem. *arXiv preprint arXiv:1907.12860*, 2019.

[56] Symantec. The need for threat risk levels in secure web gateways. https://docs.broadcom.com/doc/need-for-threat-risk-levels-in-secure-web-gateways-en, 2017.

[57] Symantec. Webpulse. https://www.symantec.com/content/dam/symantec/docs/white-papers/webpulse-en.pdf, 2017.

[58] Sarah Szczypinski. What is financial profiling? https://www.lexingtonlaw.com/blog/finance/financial-profiling.html, 2021.

[59] Izabela Tarlowska and Aleksandra Zebrowska. Customer profiling for credit decisions made easy for the financial industry under new polish legislation. https://www.jdsupra.com/legalnews/customer-profiling-for-credit-decisions-85940/, 2019.

[60] The World Wide Web Consortium. Same origin policy. https://www.w3.org/Security/wiki, 2020.

[61] Narseo Vallina-Rodriguez, Jay Shah, Alessandro Finamore, Yan Grunenberger, Konstantina Papagiannaki, Hamed Haddadi, and Jon Crowcroft. Breaking for commercials: characterizing mobile advertising. In *Internet Measurement Conference (IMC)*, 2012.

[62] Seppe Vanden Broucke and Bart Baesens. *Practical Web scraping for data science*. Springer, 2018.

[63] Thomas Vissers, Nick Nikiforakis, Nataliia Bielova, and Wouter Joosen. Crying wolf? on the price discrimination of online airline tickets. In *7th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2014)*, 2014.

[64] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L Mazurek, and Blase Ur. Oh, the places you've been! user reactions to longitudinal transparency about third-party web tracking and inferencing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 149–166, 2019.

[65] Wikipedia. Filter bubble. https://en.wikipedia.org/wiki/Filter_bubble, 2021.

[66] Zenith. Adspend forecast live. http://adforecast.zenithmedia.com/, 2020.

Appendix

A Geographic location implications

Around 95.5% of the websites show identical or very similar trackers when accessed from US and a different country among France, Brazil and Australia. We verify to what extent the crawling from different geographic locations influences the veracity of our findings: we select users from one of the three countries and run all the experiments of our study, comparing the curve obtained considering trackers scanned from users' respective countries and from US.

At a glance, trends show no substantial difference, thus a similar shape regardless of the scanning location, confirming the goodness of our analysis. While exhibiting the same behavior (e.g., exponential in Figures 1, 2, 6 and 7, logarithmic in Figure 3, normal in Figure 9), we only notice a very subtle shift for some portions of the curves. Likely, the explanation is due to the fact that for websites on which we detect very different tracking organizations if scanned from different locations ($\sim$4.5%), those scanned from US have a few more trackers than the ones scanned from one of the three countries.

However, considering the restricted portion of domains for which this discrepancy exists, its implications are almost negligible: indeed, for the curve that considers trackers scanned from US, we register an average of <1 [<3] increase of new trackers encountered per i^{th} browsing hour [website] when $i<10$ [$i<5$] (Figures 1 and 2). On the contrary, we measure an average decrease of $<2\%$ when evaluating the percentage of trackers deleted when the cleaning frequency is >100 websites and >18 hours. We finally register a $<2\%$ rise when assessing the percentage of known history among the top 20 trackers.

B Users' Geographical Breakdown

Table 9: Overview of the continents and their top-2 countries ordered by percentage of users in our dataset.

Continent Country	% Users	% Trackers	Categories
North America	44.16		
United States	37.08	80.30	92
Canada	2.63	59.84	92
Mexico	2.61	43.75	91
Asia	20.87		
Philippines	6.27	58.54	93
India	3.97	51.16	92
Malaysia	2.70	40.16	88
Europe	18.69		
Great Britain	4.10	66.33	91
France	2.35	51.17	90
Italy	1.72	46.71	88
South America	9.33		
Peru	2.05	43.83	91
Brazil	1.95	39.40	90
Colombia	1.91	38.62	85
Africa	5.00		
Nigeria	1.59	30.08	83
South Africa	1.15	39.76	89
Egypt	0.66	34.05	91
Oceania	2.74		
Australia	2.12	53.34	91
New Zealand	0.44	37.83	90
Fiji	0.10	22.37	75

C Website categorization

Website categories are provided by a public classification service from the security vendor [57]. The service supports over 60 languages and is composed of more than 300 specialized modules that disassemble web pages and analyze their components. The main features used to feed the classification algorithm are: webpage language, source code language, document type, character set, external link categories, content words, scripts and iframes. In addition, the categorization is fine-tuned by an offline system, which simultaneously analyzes multiple pages looking for connections and additional evidence to supplement what was collected in real time. HTTP referrer headers and hyperlinks are examples of attributes used in this phase.

D Website security risk scores

Website security risk scores are obtained by querying a public risk-level calculator service offered by the security company [56]. The service uses cloud-based artificial intelligence (AI) engines to categorize websites by combining multiple data sources. At first, historical information of the domain is used to detect the existence of malicious behaviors, e.g., whether its DNS resolutions belong to malicious networks and the website has already been identified as source of malware, scams or phishing. The webpage is then queried and the characteristics of its content together with features extracted from the server behavior are analyzed (e.g., shady file content, network errors, lie detector analysis). The AI algorithm then outputs a risk score between 1 and 10, going from websites with huge traffic and long history of good behavior (risk 1), through webpages with evidence of shady behavior (risk 5), to domains with solid evidence of maliciousness (risk 10).

E Users' browsing days and hours distribution

Figure 13, summarizes the users' activity in terms of mean browsing days and hours For each user, we report the number of days and hours per day in which we observe at least one record in the telemetry, and report the distribution of the whole dataset in the form of a boxplot.

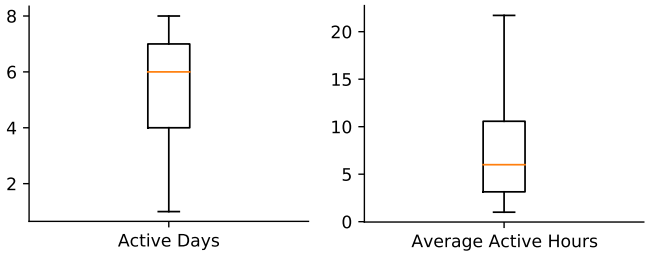


Figure 13: Overview of the average number of active days and hours per day for all the users in our dataset. Orange horizontal lines represent medians for each group. The lower and upper interquartile ranges are respectively the 25th and 75th percentile.