The Internet with Privacy Policies: Measuring The Web Upon Consent

NIKHIL JHA, MARTINO TREVISAN, LUCA VASSIO, MARCO MELLIA, Politecnico di Torino, Italy

To protect users' privacy, legislators have regulated the usage of tracking technologies, mandating the acquisition of users' consent before collecting data. Consequently, websites started showing more and more consent management modules – i.e., Privacy Banners – the visitors have to interact with to access the website content. They challenge the automatic collection of Web measurements, primarily to monitor the extensiveness of tracking technologies but also to measure Web performance in the wild. Privacy Banners in fact limit crawlers from observing the actual website content.

In this paper, we present a thorough measurement campaign focusing on popular websites in Europe and the US, visiting both landing and internal pages from different countries around the world. We engineer *Priv-Accept*, a Web crawler able to accept the privacy policies, as most users would do in practice. This let us compare how webpages change before and after. Our results show that all measurements performed not dealing with the Privacy Banners offer a very biased and partial view of the Web. After accepting the privacy policies, we observe an increase of up to 70 trackers, which in turn slows down the webpage load time by a factor of 2x-3x.

Additional Key Words and Phrases: Web Measurements, Crawling, Privacy Banner, GDPR

1 INTRODUCTION

The Web is a complex ecosystem where websites monetize their audience through advertising and data collection. They use Web trackers, i.e., third-party services that collect the visitors browsing history, to build per-user profiles and display targeted ads and personalised content [13, 41, 44]. Hundreds of tracking platforms exist, with many of them gathering information from a large base of users and websites [31, 36, 39, 43].

This picture has created tension over users' online privacy, and regulatory bodies have started governing the scenario. Lastly, in May 2018, the EU introduced the General Data Protection Regulation (GDPR) [30]. It sets strict rules on collecting and storing personal data and mandates firms to ask for informed consent. Similarly, the California Consumer Privacy Act of 2018 (CCPA) [20] gives consumers more control over the personal information that businesses collect. All this has changed the Web too. Nowadays, when users visit a website for the first time, a consent management module – the commonly called Privacy Banner – prompts, asking the visitors whether they accept the website privacy policy and the use of tracking techniques, and eventually which tracking mechanisms to accept or to block. Upon user's acceptance, the browser activates the accepted tracking techniques and updates the webpage to include all ads and third-party objects.

This challenges the commonly accepted approach to automatically crawl websites to measure the Web ecosystem on privacy [13, 28, 31, 35, 36, 38, 39, 41, 43, 43, 44, 49, 51] and performance [14, 16, 19, 25, 29, 40, 45, 47, 55]. These measurements are typically carried out with headless browsers to access webpages of popular websites and to automatize the collection of metadata and statistics. However, today, these measurements could result biased and unrealistic, with the crawler observing possibly very different content than what a user would get after accepting the privacy policies – as most users would commonly do [18, 27, 33]. While researchers have shown the importance of carefully choosing which webpages to test [15], to the best of our knowledge, we are the first to consider the impact of Privacy Banners on automatic measurements.

Author's address: Nikhil Jha, Martino Trevisan, Luca Vassio, Marco Mellia, Politecnico di Torino, Italy, first.last@polito.it.

For this, we engineer *Priv-Accept*, a tool to automatically handle the privacy acceptance mechanisms the websites put in place. In a nutshell, *Priv-Accept* enables the collection of user-like Web measurements. It overcomes the limitations of traditional crawling approaches, allowing the measurement of the tracking ecosystem to which users are actually exposed and obtain thus realistic figures on performance. The non-standard way of displaying the Privacy Banner, the presence of multiple languages, and the freedom to customize the accept button make automatic detection and acceptance not trivial. We base *Priv-Accept* on a keyword list that we thoroughly build to accept the privacy policies automatically. Compared to other solutions [3, 7–9], *Priv-Accept* proves the most robust approach, bypassing the Privacy Banner in about 90% of cases when present.

Armed with *Priv-Accept*, we run an extensive measurement campaign. We focus mostly on European and US websites that we visit from different countries. We demonstrate how different is the picture we observe before and after accepting the website privacy policies. Interestingly, many websites correctly implement the regulations, and they activate trackers and personalized ads only after consent is collected. This makes the illusion that tracking is decreasing with respect to the past [35]. However, the number of trackers websites embed substantially increases upon acceptance of the privacy policy, in some cases up to 70. As such, popular trackers suddenly become much more pervasive than one can measure using traditional and naive Web crawlers. Considering performance, after accepting privacy policies, webpages become more than three times heavier and more complex, loading objects from many more third-party websites. Thus, they are slower to load, so that webpages embedding many trackers and ads double or triple the webpage load time.

Recently, authors of [15] showed how important it is to extend the crawling to internal pages. Here, we show that it is on par fundamental to correctly handle the Privacy Banners when running extensive Web measurements. For this, we offer *Priv-Accept* as an open-source tool to incentive also other researchers to contribute to it. Similarly, we offer all the data we collected for this study to the community in an effort to support reproducibility and foster other studies.[1]

After discussing the scenario and related work in Section 2, we present *Priv-Accept* and thoroughly test it in Section 3. In Section 4, we report how different the picture results when checking the Web tracking ecosystem before and after the acceptance of the privacy policies. We then show the implications on performance in Section 5. After discussing Ethics in Section 6, we summarize our findings in Section 7.

2 BACKGROUND AND RELATED WORK

Content providers on the Web (websites, social networks, etc.) often monetize the content they offer using advertisements. To increase their effectiveness, the so-called behavioral advertisement leverages users' interests to provide targeted ads. This is possible thanks to Web trackers, i.e., third-party services embedded in the webpages that gather users' browsing history. Trackers are nowadays largely present on websites and reach the majority of internauts [39, 43]. Trackers exploit cookies and advanced techniques to enable the collection of personal information [13, 41, 44].

2.1 The Role of Legislators

In this tangled picture, legislators started to regulate the ecosystem to avoid massive indiscriminate tracking that may threaten users' privacy. The first attempt has been the European Cookie Law [21] entered into force in 2013, which mandates websites to ask for informed consent before using any profiling technology. In May 2018, the General Data Protection Regulation (GDPR) [30] entered into force in all European member states. It is an extensive regulation on privacy, aiming at protecting users' privacy by imposing strict rules when handling personal information. Unlike previous

[1] *Priv-Accept* is available as an open-source GitHub project at: https://github.com/marty90/priv-accept

Fig. 1. Example of Privacy Banner on `dailymail.co.uk`. Only upon consent, trackers are contacted and ads displayed.

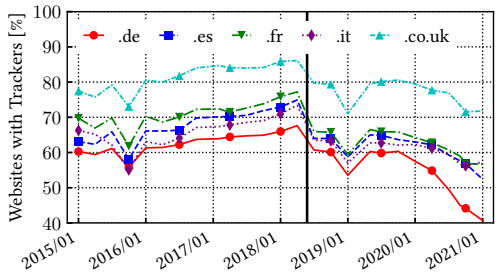


Fig. 2. Percentage of websites containing at least one tracker (from HTTPArchive). The black vertical line indicates the entry into force of the GDPR.

regulations, it sets severe fines and infringements that could result in a fine of up to €10 million, or 2% of the firm's worldwide annual revenue, whichever amount is higher. Some websites have already been caught to present legal violations in their Cookie Banner implementation [37] and a large fraction have been shown to use tracking technologies before users consent [46, 50]. In the US, the California Consumer Privacy Act (CCPA) [20] similarly enhances privacy rights and consumer protection for California residents by requiring businesses to give consumers notices about their privacy practices.

As a result, most of the websites now provide explicit Privacy Banners [26] and many adopt Consent Management toolsets [34], making the website content difficult to access until visitors accept the privacy policy. For example, Figure 1 shows the same news website homepage before and after accepting the privacy policy. Only upon pressing the "Got it" button, the website content is fully visible.

2.2 The Effect of Privacy Banners on Web Measurements

Despite cases of misuse, the new regulations had a large impact on the internauts, and this complicates the measurement of the tracking ecosystem. A simple Web crawler visiting the websites without accepting the privacy policies would offer a biased picture, with no trackers and no ads being loaded. Hu *et al.* [35] already found that the number of third-parties dropped by more than 10% after GDPR when visiting websites automatically. Conversely, when using a dataset from 15

real users, they measure no significant reduction in long-term numbers of third-party cookies. Dabrowski *et al.* [24] draw similar conclusions, finding an apparent decrease in the use of persistent cookies from 2016 to 2018. Sorensen *et al.* [48] testify a decreasing trend in the number of third parties during 2018. We quantify this phenomenon in Figure 2, using the HTTPArchive open dataset [6]. The curators of this dataset maintain a list of top websites worldwide that they automatically visit using the Google Chrome browser from a US-based server to store a copy of each visited webpage. Using the tracker list detailed in Section 3, we report the percentage of websites embedding one or more trackers for 5 European countries (simply using the Top-Level Domain to identify the country). We restrict the analysis on those websites that exist for the whole six years-long periods (9 196 website in total).

Figure 2 apparently shows that the introduction of the GDPR (the black vertical line in May 2018) results in an abrupt decrease in the number of tracker-embedding websites, a trend that continues up to the moment we write. However, as we will show, these measurements are an artifact due to the GDPR itself. Indeed, the Web crawler used by HTTPArchive can only capture the behavior of the websites as a "first-time visitor", before the user accepts any privacy policy. The crawler thus misses cookies, third-party trackers, and any personalized ads.

Research papers that rely on crawling large portions of the Web for different reasons could be affected by the same bias in their measurements. For instance, this would challenge the automatic measurement of the Web ecosystem on privacy [13, 15, 28, 31, 35, 36, 39, 41, 43, 44, 51] and countermeasurements [38, 43, 49]. Moreover, this will also impact those works that rely on crawlers and headless browsers [17] to quantify the impact in the wild of new technologies like SPDY, HTTP/2 [19, 25, 29, 55], 4G/5G [14, 16], accelerating proxies [45, 47, 56], or generic benchmark solutions [40]. At last, even spiders and mirroring tools like HTTPArchive may be affected if the website allows the visitor to access its content only after accepting the privacy policy.

2.3 Related Work and Tools

Authors of [52] are the first to consider the impact of the Privacy Banner presence. First, they instruct a custom OpenWPM crawler to identify specific Cookie Banners, and then they manually verify the results. Unfortunately, they solely focus on the pornographic ecosystem, which they acknowledge to be rather different from the Web at large, and thus their work can hardly be generalized.

Recently, authors of [15] demonstrated that it is fundamental to consider the complexity of the Web ecosystem and include internal pages in every measurement study. They find a number of recent works that neglect internal pages and, as such, might provide biased results. Yet, they ignore the implications of Privacy Banners. Here, we aim at providing an extensive and thorough study of their impact on the Web. Our goal is to enable the study of webpage characteristics as visitors would experience, assuming that most of them accept the default privacy setting as offered by the Privacy Banner. Indeed, it has been shown that users tend to ignore privacy-related notices [22, 32, 54]. Considering GDPR Privacy Banners, users tend to accept privacy policies when offered a default button via intrusive banners that nudge users [18, 27], which is often the case [33] with websites offering large pop-ups or wall-style banners that cover most of the webpage as seen in Figure 1.

There exist solutions that aim at automatically managing Privacy Banners: some browser add-ons try to hide Privacy Banners using a list of CSS selectors of known Privacy Banners. The most popular add-ons of this kind are "I don't care about cookies" [7] and "Remove Cookie Banners" [9]. Unfortunately, hiding the Privacy Banners has an unpredictable behavior, in some cases falling back to privacy policies acceptance, while, in other cases, triggering an opt-out choice. Other proposals, again in the form of browser add-ons, try to explicitly opt-in or opt-out to cookies. For example, "Ninja Cookie" [8] approves only cookies strictly needed to proceed on the website. Conversely,

Autoconsent [2] and Consent-O-Matic [3] use a set of predefined rules to either opt-in or opt-out to cookies, according to the user configuration. These two are the most similar solutions to *Priv-Accept*, as they allow to automate the action of providing consent to privacy policies if used in combination with a crawler. However, they are based on a list of actions the browser has to automatically run when finding a set of popular Consent Management Providers (CMPs), limiting their effectiveness. In Section 3.2, we compare *Priv-Accept* with Consent-O-Matic – the most mature tool – showing that it accepts privacy policies on a much smaller portion of websites than *Priv-Accept*. Indeed, the diversity of the Web ecosystem, the presence of multiple languages and the fully customizable choice of cookie banner buttons make the engineering of *Priv-Accept* not trivial.

3 *PRIV-ACCEPT* DESIGN AND TESTING

We explicitly engineer *Priv-Accept* to fully automate the visit to websites and collect statistics. The key element of *Priv-Accept* is its ability to identify the presence of a Privacy Banner and automatically accept privacy policies. We aim at a practical and effective approach to accept privacy policies through the offered button. As previously said, most users will indeed be nudged in this direction, being the opt-out options often made cumbersome on purpose[18, 27, 33].

To illustrate *Priv-Accept* operation, consider again Figure 1. A large Privacy Banner appears on the first-time-ever visit, and the user shall click on the "Got it" button to access the webpage content. *Priv-Accept* has to locate this button and click on it automatically. As a result, the website starts loading advertisements and contacting trackers in background. We refer to these two types of visits as *Before-Visit* and *After-Visit* in the remainder of the paper.

We implement *Priv-Accept* using the Selenium browser automation tool [17], the de-facto standard for browser automation. We focus on Google Chrome, but we could easily extend it to other browsers.

Given a target URL, *Priv-Accept* carries out the following tasks:

(1) It navigate to the URL with a fresh browser profile, i.e., with an empty cache and cookie storage. This makes the visit the equivalent of a *Before-Visit* to the website.
(2) It inspects the Document Object Model (DOM) of the rendered webpage to find a possible *Accept-button* in a Privacy Banner. For this, we match a list of keywords on the text of each node of the DOM. We identify an Accept-button if we exactly match one of our keywords. For robustness, the match is case insensitive, and leading, trailing or repeated blank characters are removed.
(3) If *Priv-Accept* finds the Accept-button, it tries to accept the default privacy policies by clicking on the corresponding DOM element (typically a `<button>`, `<href>` or `<span>` element).
(4) *Priv-Accept* then revisits the URL to collect statistics about the *After-Visit* experience.

In the beginning, we built *Priv-Accept* to look for accept buttons through CSS selectors combined with keywords as done in [52] and popular add-ons. However, we soon observed that this methodology was too fragile as the use of selectors is strongly CMP-specific and highly customizable by webmasters. The keyword-based approach eases the generalization of the solution. Considering the complexity, *Priv-Accept* adds marginal overhead to the time required to visit a webpage. Only for very complex webpages, iterating through all DOM elements may require some time, but this is still much less than the time needed to load and render the webpage by the browser.

During each visit, *Priv-Accept* stores metadata regarding the whole process in a JSON log file. It includes details on all HTTP transactions and installed cookies. Moreover, it optionally takes screenshots of the webpage during the various phases to allow manual verification.

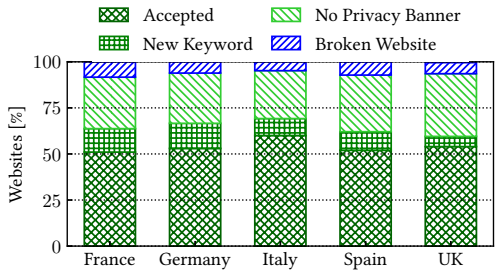


Fig. 3. Validation results of *Priv-Accept* over 200 randomly picked websites per country.

Priv-Accept is highly customizable and offers the user various features. It lets the user customize the declared `User-Agent` and browser language (in the `Accept-Language` headers). Important to our analysis, it runs a:

- *Warm-up visit*: to populate the browser cache.
- *Before-Visit*: to collect statistics on the webpage before accepting the privacy policy, as a Naive Crawler would do.
- *After-Visit*: to collect statistics on the webpage as it appears after accepting the privacy policy (if an Accept-button is found).
- *Additional-Visits*: to a number of webpages of the same website, randomly choosing among the internal links. This step runs regardless of the presence of the *Accept-button*.

Among metadata *Priv-Accept* collects, we record the Page Load Time, or *OnLoad* time, on all visits. It allows us to compare the performance with and without privacy acceptance. The *OnLoad* time is a performance index often used as a proxy for Quality of Experience measurements [23]. We leave the measurements of more sophisticated QoE-related metrics such as the SpeedIndex [11] as future work. Moreover, we neglect metrics that are not affected by the presence of a Privacy Banner, such as the Time-to-first-byte (TTFB). To avoid suffering the bias of the *After-Visit* that can only occur with a warm browser cache, we run a preliminary *Warm-up visit*, then we perform another *Before-Visit* and take performance measurement only on the latter. This lets us fairly compare the *OnLoad* on the two visits with hot cache in both cases. Alternatively, *Priv-Accept* can erase the HTTP cache and clean the socket pool upon each visit to measure webpage performance with a cold cache.

At last, to limit the impact of random delay due to webpage download and rendering, *Priv-Accept* uses quite conservative timeouts before eventually abort the visit. In detail, the DOM inspection starts 5 seconds after the *OnLoad* event. While this clearly slows down the visit of multiple webpages, it maximizes the accept success rate.

To allow large-scale measurement campaigns, we containerize *Priv-Accept* using the Docker container engine [4]. In the containerized version, we use Google Chrome version 89 in headless mode and force it to use a standard `User-Agent` instead of the pre-defined `ChromeHeadless`.

We offer *Priv-Accept* as open-source to foster its usage and allow the reproducibility of the results presented in this paper. For this, we also commit to releasing all the data we collected for this study.

3.1 Keyword Selection and Validation

The core of *Priv-Accept* is the list of keywords to be matched against the webpage content to localize the clickable DOM element for accepting the privacy policy. We thoroughly build this list manually

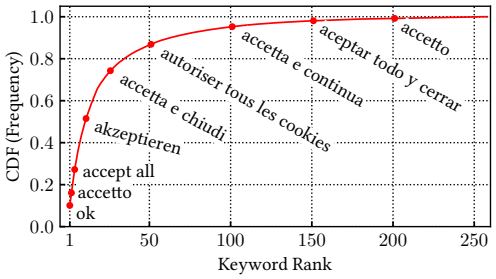


Fig. 4. Frequency of the *Priv-Accept* keywords, with some examples reported.

in an iterative way. To handle different languages, we build a list that includes keywords for each country we are interested in. For this work, we focus on 5 European countries, namely France, Germany, Italy, Spain, UK[2], plus the US – which we use as an example of an extra-EU country. For each country, we pick the most popular websites according to the Similarweb lists [10], a website-ranking service analogous to Alexa.

3.1.1 First Round - keyword extraction from top websites. In the first round, for each of the 5 countries, we consider the top-200 websites that have a Privacy Banner. We randomly choose half of these websites and manually visit them (from Europe) to extract the accept keyword. In total, we identify 186 unique keywords. We next instruct *Priv-Accept* to visit the other half of websites and accept privacy policies. For those where it fails, we manually visit them and extract keywords. With this, we include 36 new keywords, 222 in total.

3.1.2 Second Round - testing and keyword increase. To evaluate the accuracy of *Priv-Accept* in the wild, we next consider 200 new random websites for each country from the Similarweb lists. We let *Priv-Accept* visit them and manually check the subset of websites for which *Priv-Accept* fails to accept the privacy policy. We depict the results in Figure 3. *Priv-Accept* can accept the privacy policy in more than half of websites. In 6 – 14% of cases, we find new keywords – that we promptly add to our list. Interestingly, we find a non-negligible portion of websites (26 – 30%) that do not present any Privacy Banner. At last, *Priv-Accept* fails to accept privacy in only 5 – 8% of cases. Investigating further, this is due to non-standard behaviors of the webpage when accessed in headless mode. For instance, some websites present a CAPTCHA when they detect an automated visit; other websites return a blank webpage. This is a common problem for any crawler-based measurement study [53]. Note that cases of *False Positives* – i.e., *Priv-Accept* clicking on a wrong DOM element – are possible, although we have not observed any during the development and testing phases.

At the end of the keyword list building phases, we collect a total of 258 keywords covering 6 languages.[3] The most frequent one is the simple "Ok" string. In Figure 4 we show the cumulative distribution of keyword frequency on the whole set of Similarweb websites with some keyword examples. The top-50 keywords already cover 87% of websites, while 100 are enough to cover 95%. Interestingly, we find also complex strings like "I'm fine with this" or "Alle auswählen, weiterlesen und unsere arbeit unterstützen".[4]

[2]Since January 2021 UK has enforced the UK GDPR - with practically identical requirements.
[3]In Spain, some websites are in Catalan, other than in Spanish.
[4]Which translates to "Select all, keep reading and support our work".

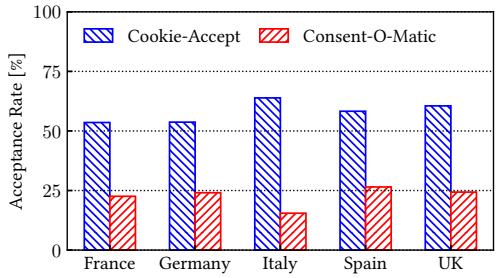


Fig. 5. Privacy policy acceptance rate of *Priv-Accept* and Consent-O-Matic on 100 websites per country.

3.2 *Priv-Accept* vs. Consent-O-Matic

We compare the effectiveness of *Priv-Accept* with Consent-O-Matic, the most mature browser plugin designed to offer/deny consent to privacy policies automatically. Unlike our tool, Consent-O-Matic exploits the presence of popular Consent Management Providers (CMP), services that take care of the management of users' choices on behalf of the website. At the time of writing, Consent-O-Matic allows managing Privacy Banners for 35 CMPs. To gauge its performance, we visit the top-100 most popular websites with a Privacy Banner for the 5 countries using a Chrome browser with the Consent-O-Matic plugin enabled. Consent-O-Matic accepts the privacy policies in less than 35% of websites with Privacy Banner, and as little as 17% and 20% for websites in Italy and UK, respectively. Here *Priv-Accept* accepts the privacy policies on all websites by construction.

We then run a second experiment considering another set of 100 websites randomly picked from the Similarweb per country lists. We visit each website with *Priv-Accept* and a Consent-O-Matic-enabled browser. Figure 5 summarizes the comparison. *Priv-Accept* can accept the privacy policies in more than 50% of websites, more than twice the success rate of Consent-O-Matic. These results are in line with those of Figure 3. The remaining websites may not have a Privacy Banner, fail to load, or use an unknown keyword. This testifies that the customization of Privacy Banners makes it difficult to engineer a generic and simple solution. The keyword-based strategy results more robust than the CMP-based approach (with similar complexity in curating the lists).

3.3 Dataset and Tracker list

In the following, we use *Priv-Accept* to run several measurement campaigns. Most of our analyses, unless otherwise indicated, targets a large set of websites popular in Europe, using a test server located in a European university campus. We also use the US as a representative of an extra-EU country. For each of the 6 countries, we consider the top 100 websites from 25 different categories - see Figure 10. In total, we include 12 277 unique websites to visit (as the lists in different countries partially overlap).

We run *Priv-Accept* on the target websites using a single high-end server running 16 parallel instances to speed up the crawl. We instrument it to run a *test sequence*, which consists in a *Warm-up visit*, *Before-Visit*, and *After-Visit* to the landing page, followed by *Additional-Visits* to 5 randomly chosen internal pages – previous studies indeed show that internal and landing pages have different properties [15]. For each website, we repeat the test sequence 5 times, randomizing the order of websites to visit in each repetition. Our main experimental campaign took place for two weeks on April 2021.

We run additional measurement campaigns to investigate specific aspects. First, we repeat the above experiments using servers located in the US, Brazil and Japan. We use Amazon Web Services to deploy on-demand servers on the desired availability zone. Our goal is to understand whether Privacy Banners appear or have a different impact depending on the visitor location. Second, we visit the top-100 000 websites according to the Tranco list [42] to offer a view on a larger number of websites. Unfortunately, the Tranco list does not offer a per-category and per-country rank. We test these websites twice, with and without clearing the browser cache, to compare webpage performance on the *Before-Visit* and on the *After-Visit* both with a warm and a cold cache.

To observe how the presence of trackers changes from the *Before-Visit* to the *After-Visit*, we rely on publicly-available lists provided by Whotracksme [12] (a tracking-related open-data provider), EasyPrivacy [5] (one of the lists at the core of AdBlock tracker-blocking strategy) and AdGuard [1] (another ad-blocking tool). For robustness, we merge the three lists and consider as potential trackers those third-party domains that appear in at least two lists. In total, we obtain 1 497 domains that we consider tracking services.[5] We then record the presence of a tracker during a visit if the webpage embeds an object from a tracking domain, and the latter installs a cookie with a lifetime longer than one month [50] – commonly referred to as *profiling cookie*. As such, we divide the HTTP transactions carried out during a visit in:

- First-Party: objects from the same domain of the target webpage.
- Third-Party: objects from a different domain than the target webpage.
- Trackers: objects from a Third-Party that is a tracking domain and sets a profiling cookie.

4 IMPACT ON TRACKING

In this section, we characterize how the Web tracking ecosystem changes if observed with or without accepting the privacy policies. We break down results by Third-Party/Tracker, by country and website category. For this, we focus on the list of 100-top popular websites per country and category. Out of the 12 277 websites, *Priv-Accept* accepts the privacy policies on 57.3% of them. The percentage is not uniform across countries and it is generally higher on European ($59-65\%$) and lower for US (35%) websites - despite most keywords are in English. Differences are more pronounced across categories; *Priv-Accept* accepts privacy policies on 87% of News websites, while only on 20% of Adult portals. For the sake of completeness, the per-country and per-category acceptance rate is reported on the top-x labels of Figure 9a and 10a, respectively. These figures are in line with the acceptance rates seen in Figures 3 and 5. Some manual random checking confirms that *Priv-Accept* does not accept the privacy policy for those websites that do not present any Privacy Banner or where the headless browser fails to visit.

4.1 Third-Party and Tracker Pervasiveness

We first study the pervasiveness of Third-Parties and Trackers and check how it varies when we measure it in a *Before-Visit* or *After-Visit*. We here focus on the 10 542 websites that are popular in the European countries according to the Similarweb ranks. Indeed, we aim at quantifying the impact of privacy policy acceptance on European websites. As such, we exclude those websites exclusively popular in the US.

We first detail the top-15 most pervasive Third-Parties in Figure 6. The GDPR mandates to obtain informed consent before starting to collect any personal data. As such, Third-Parties may be seen as possibly offending services if activated before accepting the privacy policy.[6] With little

[5] In the following, we identify them with their *second-level domain name* – i.e., a hostname truncated after the second label. We handle the case of two-label country code TLDs such as `co.uk`.

[6] Here, we do not enter into the debate of what can be considered a Tracker.

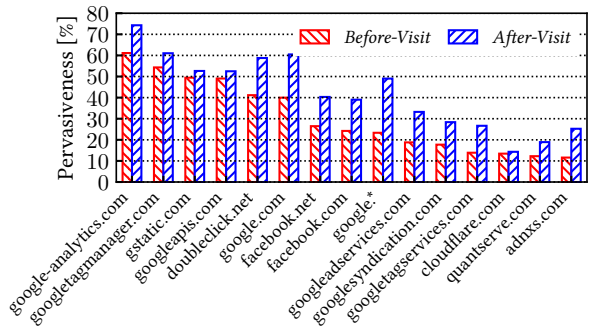


Fig. 6. Pervasiveness of the top-15 Third-Parties.

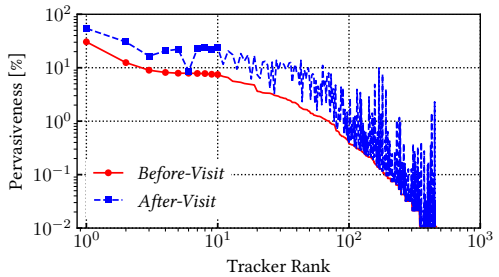


Fig. 7. Pervasiveness of the 342 identified Trackers.

surprise, the most pervasive Third-Party is `google-analytics.com`. It grows from 61% to 74% in popularity on the *After-Visit*. The growth is also sizeable for other Google services such as `googleadservices.com` and `googlesyndication.com`. Conversely, domains belonging to Content Delivery Networks, such as `cloudflare.com` and `cloudflare.net` do not increase their pervasiveness on the *After-Visit*, likely being not included in the mechanisms of Privacy Banners. Interestingly, only 3 out of the top-15 Third-Parties are Trackers – i.e., present in our tracker list and setting a persistent cookie. `doubleclick.net` and `facebook.com` are the most popular ones, with pervasiveness growing from 41% to 58% and from 24% to 39% on the *After-Visit*, respectively. They are present in more than twice the number of websites than their first competitor (`quantserve.com`).

Focusing now on Trackers only, Figure 7 shows their pervasiveness in our dataset. We count 342 of them. Notice that the figure has log-log axes to better show the large variability of Tracker popularity. The red curve shows the pervasiveness on the *Before-Visit*, which is what a naive crawler would report. The blue curve shows how the figure changes on the *After-Visit*. The increase in pervasiveness is general and includes both popular and infrequent Trackers, reaching in a few cases one order of magnitude. On the *After-Visit*, the number of Trackers that are present on 1% or more of websites grows from 40 to 90. Interestingly, if we sort the *Before-Visit* and *After-Visit* Trackers by their pervasiveness, the rank remains almost unchanged. The Spearman's rank correlation is 0.90, indicating that the Tracker popularity order is approximately the same before and after the privacy policy acceptance. The difference is that their presence increases.

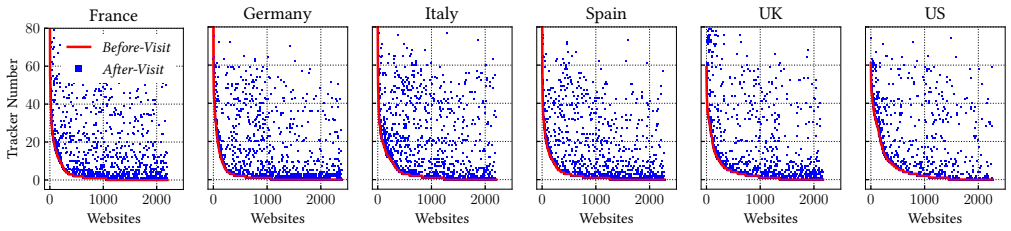
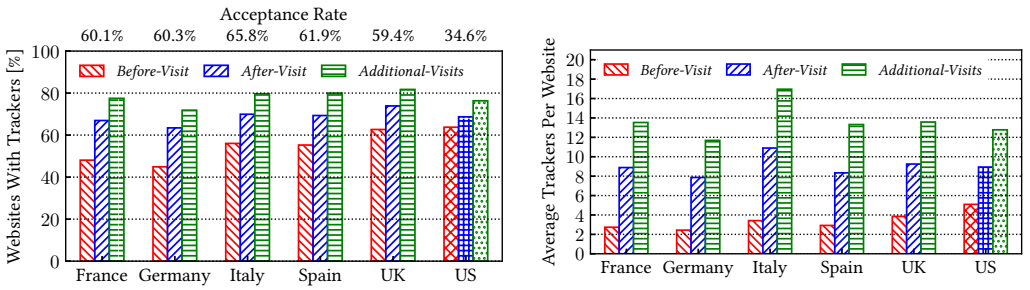


Fig. 8. Trackers per website seen on the landing page. Websites are sorted by Tracker number on the *Before-Visit*.



(a) Percentage of websites embedding Trackers.

(b) Average number of Trackers per website.

Fig. 9. Trackers penetration and number on websites during different phases of a browsing sessions.

As it emerges from Figure 7, many Trackers are widespread even on the *Before-Visit*. This hints at a possibly wrong implementation of the GPDR regulation, which mandates to acquire first the visitor's explicit consent before activating any tracking mechanisms. To be precise, the presence of Trackers on the *Before-Visit* does not necessarily entail a violation of the law. A manual analysis displays that some Trackers install test cookies during the *Before-Visit* using a form similar to `test_cookie = CheckForPermission`. These cookies are just a check for the possibility of installing profiling cookies upon the user's acceptance. It is thus possible that the *Before-Visit* pervasiveness of some Trackers includes cases in which only test cookies are actually used. Here we limit to observe that often Trackers set some (potentially) profiling cookies even on the *Before-Visit*.

In conclusion, these results show how different the picture is when collecting measurements with or without accepting the privacy policies. *Priv-Accept* enables the collection and analysis of what most users would be exposed to, thanks to its ability to handle the Privacy Banners and accept website privacy policies.

4.2 Breakdown on Websites

We now detail the impact of accepting privacy policies on the number of Trackers found in each website, breaking down our results by country and website category.

4.2.1 Analysis by country. We first check if the number of Trackers found during the *Before-Visit* differs in the *After-Visit*. Figure 8 shows websites sorted in descending order by the number of contacted Trackers as measured in the *Before-Visit* (red curve). Focus now on the blue points. They report the number of Trackers in the *After-Visit* for the same website. The number tends to grow on the *After-Visit*, underlying again the need for tools such as *Priv-Accept* to accept the privacy

policy and measure the footprint of Web tracking correctly. Some websites present a sizeable increase, with figures that grow by 50-70 Trackers. Curiously, some websites that already include Trackers in the *Before-Visit* include more Trackers in the *After-Visit*. This possibly hints at a wrong implementation of the Privacy Banner, which fails to hinder the presence of possibly offending Trackers. The increase is less remarkable for US-popular websites – again, mainly due to the less widespread presence of Cookie Banners.

To better quantify Tracker presence, we show the fraction of websites containing at least one Tracker in Figure 9a. About 50% of websites popular in European countries already include some Trackers on *Before-Visit*. This happens more frequently in the UK (63%) and more occasionally in Germany (44%). Again, notice that a website embedding a Tracker on the *Before-Visit* does not necessarily represent a violation of the GDPR, even if this can often be the case [50]. Interestingly, in the US this figure is higher than in European countries. Recalling that in the US the probability of encountering a Privacy Banner is lower, this hints at a positive effect of the GDPR on popular European websites. The percentage of websites containing Trackers in the *After-Visit* grows for all European countries from a +11% increase in the UK to +20% for Germany. This increase is moderate (+5%) in the US, given the lower fraction of those websites having a Privacy Banner. We complete this analysis by reporting how this fraction increases when performing *Additional-Visits* as recommended in [15]. We perform 5 *Additional-Visits* per website. Our results confirm this, with the chance to observe at least one Tracker that further grows by 5%-10% in *Additional-Visits* when compared to the *After-Visit*.

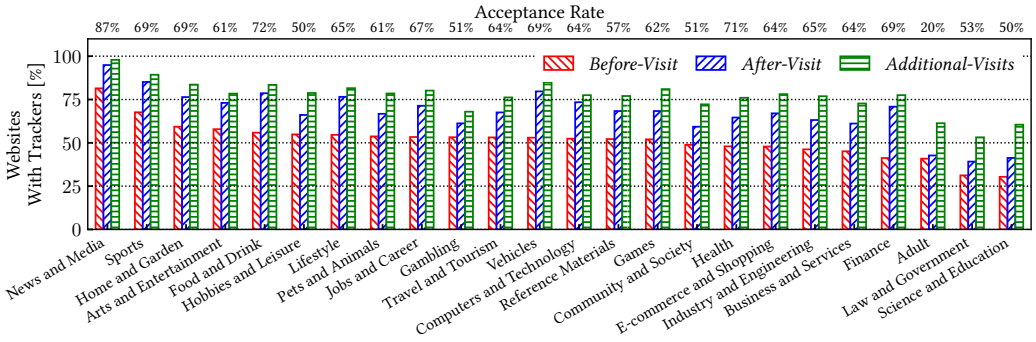
We next investigate the quantity of Trackers contacted while visiting websites in Figure 9b, which shows the average number of Trackers contacted on the websites, separately by country. For all countries, the average amount of Trackers more than doubles on *After-Visit*, and performing *Additional-Visits* further increases this figure. In Italy, for instance, this figure grows by a factor of 4 when comparing *Before-Visit* and *Additional-Visits*. As previously noted, the behavior of US-popular websites differs from the European: before acceptance, the number of Trackers is already higher than in popular European websites, while it is comparable after.[7] This hints that popular websites in the United States may not have to deal with the European legislation, thus being less receptive to GDPR indications. On the opposite side, German-popular websites appear to be the most observant of the regulations, installing Trackers only upon accepting the privacy policies. Afterward, they reach levels comparable to the other countries. In summary, European websites use the same quantity of Trackers as US ones, although they are often contacted only after accepting the privacy policy.

We finally observe that the probabilistic nature of web tracking and bidding mechanisms result in a different number of Trackers contacted at each visit. To obtain the most reliable measurements, we test each website 5 times. We notice that measuring the fraction of websites containing at least one Tracker (as in Figure 9a) is moderately impacted by the number of tests. Indeed, when considering a single *After-Visit* per website, overall, we find 69.1% of them containing one (or more) Trackers, which increases only to 70.0% considering all 5 visits. Similarly, the average number of Trackers (as in Figure 9b), increases from 6.5 to 7.8.

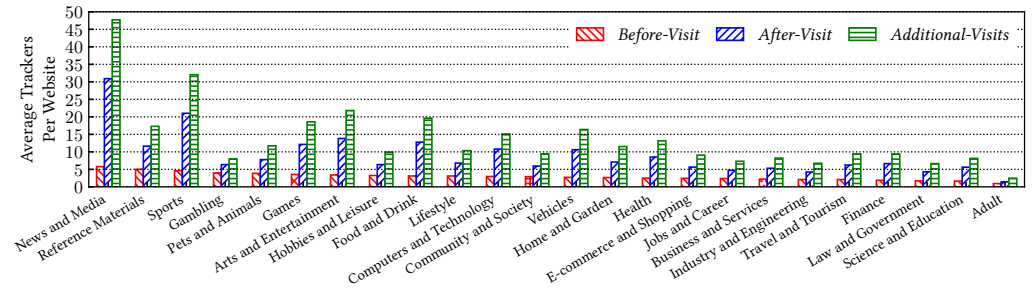
4.2.2 Analysis by category. We now break down the picture by category, showing the results in Figure 10. As we previously pointed, we explicitly target websites from 25 categories, each holding the top-100 websites for each of the considered countries.[8]

Figure 10a reports the percentage of websites of a given category that contain at least one Tracker. We sort categories from the highest to the lowest percentage of websites with Trackers in

[7] Recall that these measurements are taken from a European country.
[8] We find a handful of websites belonging to more than one category.



(a) Percentage of websites embedding Trackers.



(b) Average number of Trackers per website.

Fig. 10. Trackers penetration and number on websites during different phases of a browsing session, separately by category.

Before-Visit. For completeness, the top x-axis details the fraction of websites in such category where *Priv-Accept* accepts privacy policies. As before, there is a significant difference in the *Before-Visit* and *After-Visit*. An exception is the *Adult* category, where the increase is marginal. This is likely due to the low number of websites with Privacy Banners (20%) and confirms the peculiarity of the tracking ecosystem on Adult websites [52]. As observed in Figure 9a, considering *Additional-Visits* increases again the chance of encountering at least one Tracker.

Figure 10b shows the average number of Trackers in websites, with categories sorted from the one with average highest to the one with average lowest number of Trackers in the *Before-Visit*. In the *After-Visit* and *Additional-Visits*, there is a large increase in the number of Trackers, confirming that most Trackers appear only after the user accepts the privacy policies and when visiting internal pages. Here, differences across categories are pronounced. Categories that heavily depend on advertisement-related incomes (such as *News and Media*, *Sports*, *Games*, *Arts and Entertainment*) tend to rely on a large number of Trackers to support more effective behavioral advertisements. This is noticeable already on the *Before-Visit*. For example, access to a *News* website leads to contact 5.7 Trackers on average. Here, *Priv-Accept* successfully accepts the privacy policies in 87% of cases. Indeed, being *News* websites very popular, they tend to correctly implement the privacy regulations, showing a well-configured Privacy Banner. Upon acceptance, suddenly, the number of Trackers becomes almost 6 times higher (30.9 for *News*) and 9 times higher in *Additional-Visits*

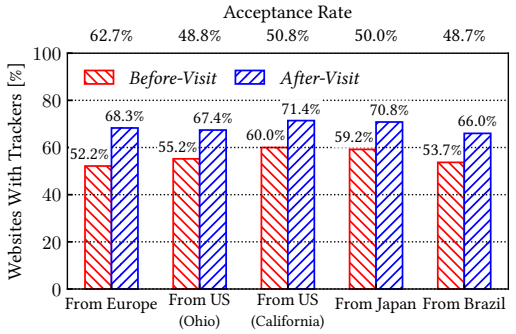


Fig. 11. Trackers per websites when crawling from different countries.

(47.7). Similar considerations hold for many other categories, e.g., for *Sport*, *Food and Drink* and *Arts and Entertainment* the average number of Trackers more than triples.

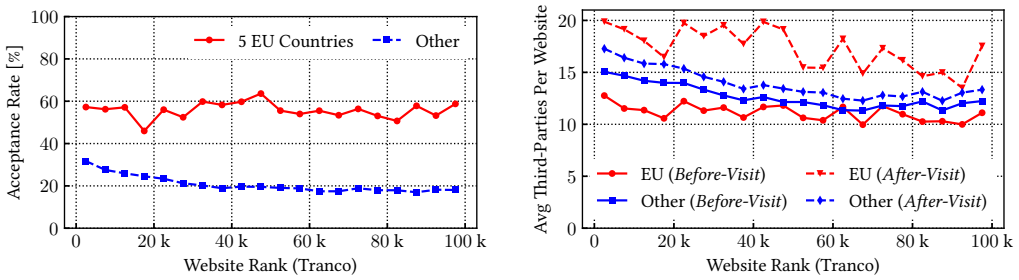
These results well highlight the need for correctly handling the Privacy Banners to observe the extensiveness of Trackers. Without *Priv-Accept*, one would radically underestimate the footprint of the tracking and ads ecosystems on the Web. In a nutshell, thanks to *Priv-Accept*, we obtain the fundamentally different figure in the *After-Visit* and *Additional-Visits*.

The case of *Adult* websites is worth a specific comment. *Priv-Accept* finds the Privacy Banner on only 20% of them, and a manual check confirms that the large majority of them do not offer any Privacy Banner. In general, tracking is also scarce upon acceptance, as previously found by authors of [52]. They suggest that the specialized pornographic advertisement ecosystem may cause this behavior: usually, trackers and advertisers related to pornographic websites do not operate outside of them – often evading tracker listing efforts.

4.3 Visiting from Outside Europe

To complete the analysis, we run additional measurement campaigns using crawling servers in the Amazon AWS data centers located in the US (Ohio and California), Japan and Brazil. We target the same websites as before. We aim to check if websites behave differently based on the location of the visitors. Figure 11 summarize our findings. First, we notice how *Priv-Accept* accepted privacy policies on around 10% fewer websites when run from outside Europe, as reported on top x-labels. Investigating further, we find $\approx 1\,150 - 1\,200$ websites for which *Priv-Accept* can accept the Privacy Banner when visited from Europe, but it fails when visited from not-EU countries. Checking the screenshot taken by *Priv-Accept* during the visit on a random subset of these websites, we confirm that no Privacy Banner is present. Thus, we conclude that some websites are starting to customize their Privacy Banners based on visitors' properties, such as their location.

This impacts the percentage of websites that embed Trackers on the *Before-Visit*. It grows from 49.7% to $54 - 60$% when visiting from outside Europe. On the *After-Visit*, these differences smooth out, revealing how *Priv-Accept* helps obtain user-centric measurements regardless of the presence or absence of Privacy Banners on websites. As a final note, we do not observe any significant difference visiting the websites from Ohio or California, despite the CCPA.[9]



(a) Acceptance rate.

(b) Average number of Third-Parties per website.

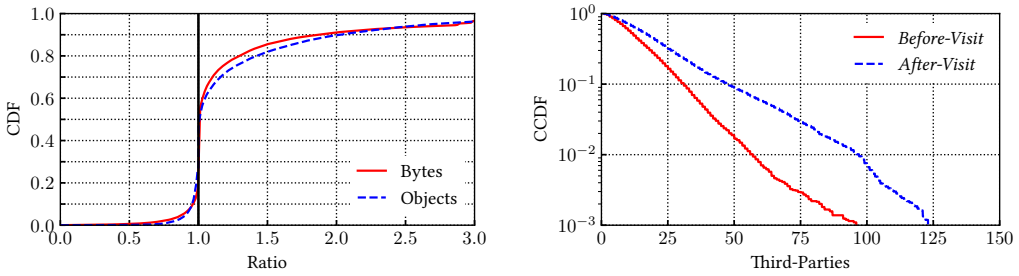
Fig. 12. Acceptance rate and average Third-Parties per website over the top-100 k websites in Tranco list, computed every 5 000 websites in the rank.

5 IMPACT ON WEBSITE COMPLEXITY AND PERFORMANCE

In this section, we measure the impact of accepting privacy policies on the webpage characteristics and loading performance. Trackers and Third-Party objects that the browser has to load and display upon consent could impact the amount of data to download and the rendering performance. Since we are not interested in breaking down results per country or per category, here we use the crawl on the top-100 000 websites according to the Tranco list. For each website, we visit only the landing page, doing a *Warm-up* visit to fill the browser cache, followed by a *Before-Visit* and *After-Visit*. We compare results on the latter two visits, considering only those websites for which *Priv-Accept* successfully accepted the privacy policy, which happens on 23% of websites. This is in line with the previous findings, as the Tranco list is a world-wide rank and includes (i) European websites in a language different from those for which we built the keyword list and (ii) websites based in non-European countries nor popular in Europe. We detail the acceptance rate on the Tranco list in Figure 12a, computed over blocks of 5 000 websites sorted by their rank, totaling 100 000 on the x-axis. The solid red line reports the acceptance rate for websites popular in the 5 European countries we target. Websites belong to this set if (i) either they appear in the Similarweb ranks for the 5 countries or (ii) the Top-Level Domain belongs to the 5 countries.[10] Out of these 6 9178 websites, *Priv-Accept* accepts the privacy policy on 3 861 (55.8%), which is close to what we have obtained with the Similarweb ranks (54.7%). Conversely, for the remaining websites (blue dashed line), the acceptance rate is 32% for the top-ranked and then it settles around 20%, hinting that some globally popular websites use a Privacy Banner even if they are based outside Europe.

The high acceptance rate for the 5 European countries results in a large increase of Third-Parties from the *Before-Visit* to the *After-Visit*, as we depict in Figure 12b, again computing values over blocks of 5 000 websites. The solid red line shows that these websites use, on average, 11.1 Third-Parties in the *Before-Visit*. In the *After-Visit*, the average grows to 17.3. Differently, the increase for the other Tranco websites is limited – see the area between the blue solid and dashed lines. In the *Before-Visit*, Third-Parties are more numerous than for the 5 European countries if we compare the solid blue and red lines. This is likely due to the larger presence of non-EU websites, not required to use a Privacy Banner. In the *After-Visit* (dashed blue line), the increase is moderate, not reaching

[9]This figure may require further investigation since we are measuring from Amazon AWS servers whose location may not be correctly handled by the CMPs.
[10]The Tranco list does not provide a per-country rank.



(a) Page size (in number of bytes and objects) ratio. (b) Number of Third Parties. Notice the log scales.

Fig. 13. Webpage characteristic before and upon consent to privacy policies (Tranco list).

the values of the 5 European countries (red dashed line), potentially because *Priv-Accept* misses many *Accept-button* in non-supported languages.

5.1 Impact on Page Objects and Size

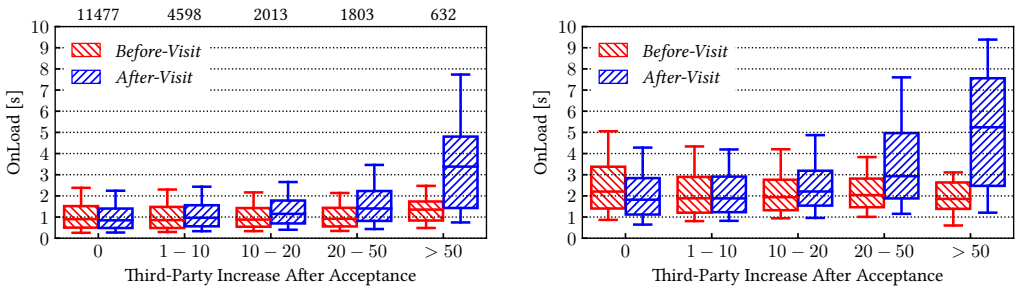
We first focus on the webpage size in terms of downloaded bytes and number of objects. To compare the results, we compute the ratio R between the measurement on the *Before-Visit* and *After-Visit*, i.e., $R = x_{After}/x_{Before}$, where x is the metric of interest. We show the results in Figure 13a, separately for total downloaded bytes (solid red line) and objects (blue dashed line). As expected, accepting the privacy policy increases the webpage size by a sizeable factor for most websites. For instance, about 9% of websites download more than twice the objects, and about 5% sees an increase of 3 times or more.

Interestingly, we also observe some websites that are lighter on the *After-Visit* than in the *Before-Visit*. Investigating further, these cases are mostly due to the lack of additional content upon acceptance and the saving of the cost of not loading the CMP objects on the *After-Visit*. This happens commonly on those websites that either add a Privacy Banner despite not using tracking mechanisms or those websites that load and contact Trackers and Third-Parties even before the user has accepted the privacy policies. While the former might be seen as an excess of caution, the latter cases are likely violating the privacy regulations.

To better characterize the differences, we quantify the number of Third-Parties seen in the *Before-Visit* and *After-Visit*. We show the Complementary Cumulative Distribution Function (CCDF) in Figure 13b. On median, websites rely on 12 Third-Parties on the *Before-Visit* (solid red line). This figure grows to 17 after (blue dashed line) on the *After-Visit*. The CCDF highlights the tail of the distribution where we observe those websites that rely on a large number of Third-Parties: the percentage of websites with more than 50 grows from 1.8% to 9.2%, with 3.0% including more than 75 Third-Parties upon acceptance.

5.2 Impact on Page Load Time

The Third-Party domains appearing after acceptance are generally devoted to advertisements, analytics and Web tracking – see in Figure 6 the most pervasive. Contacting a large number of them has direct implications on the page load time and, indirectly, on the users' QoE [23]. We thus expect the growth of Third Parties and the increase in the number of objects to download to cause degradation on the page load time since the browser has to resolve via DNS and contact many



(a) Warm Browser Cache.

(b) Cold Browser Cache.

Fig. 14. OnLoad time of websites versus the increase of Third-Party number upon acceptance (Tranco list). The cardinality of each category is reported on the top axis of the left-most figure.

servers. This possibly limits the advantages of new solutions like stream multiplexing and header compression offered by HTTP/2.

We dissect the webpage performance in Figure 14, comparing separately visits with a warm cache (Figure 14a) and a cold cache (Figure 14b). We report the *onLoad* time by grouping the website according to the different number of additional Third-Parties observed in the *After-Visit*. We use boxplots, where the boxes span from the first to the third quartile and whiskers from the 10^{th} to 90^{th} percentile. The central stroke represents the median. The number of websites in each set is detailed above the respective boxplot. The more Third-Parties are loaded upon acceptance, the larger the time needed to load the webpage and the larger its variability. Especially for the websites that add more than 10 Third-Parties, the distributions are remarkably different on the *Before-Visit* and *After-Visit*. Considering visits with cold browser cache (Figure 14a), for website with $20 - 50$ additional Third-Parties, the median *onLoad* time passes from 0.91 to 1.41 seconds. The difference increases for the 632 websites adding more than 50 Third-Parties upon acceptance. The median *onLoad* time increases from 1.35 to 3.38 seconds, more than doubling. Notice also the tail of 25% of websites loading in more than 4.8 s, which happens in less than 2% of cases during the *Before-Visit*. Similar considerations hold for visits with an empty browser cache (Figure 14b). In this case, *Priv-Accept* cleans the browser cache and the socket pool after each visit. As expected, with the clean cache, websites load generally more slowly – compare values in Figures 14a and 14b. Those that do not add new Third-Parties tend to load slightly faster on the *After-Visit*, potentially due to the absence of the Privacy Banner or CMPs. Again, we observe that those adding several Third-Parties after acceptance have larger *onLoad* time on the *After-Visit*. Indeed, the median *onLoad* time for websites adding more than 50 Third-Parties increases from 1.8 to 5.2 seconds.

In summary, measuring the webpage load time of websites without considering the implications of accepting the privacy banners would result in a very biased measurement. These results highlight how approaches such as *Priv-Accept* are fundamental to obtain a realistic picture of the Web performance and testify how actual users' experience cannot be measured without handling the Privacy Banners.

6 ETHICAL CONSIDERATIONS

During our measurements, we took care to avoid harming the crawled webpages. We contacted each website 5 times in a span of two weeks and accessed a limited number of internal webpages each time. Considering that the target of our analysis were some of the most popular websites of

Western countries, our belief is not to have caused an overload on the servers or any undesirable side effect. Moreover, since we did not interact with Third-Parties after accepting the privacy policies – included displayed ads – we consider not to have significantly altered the economic ecosystem of the crawled websites. We only used the standard HTTP and HTTPS ports for our measurements, carefully avoiding any type of port scanning procedures, and we used large timers to avoid creating any kind of congestion.

7 CONCLUSIONS

In this paper, we have demonstrated how the recent regulations have changed the Web scenario, challenging its automatic measurements through traditional Web crawlers. Websites now massively deploy Privacy Banners to obtain visitors' consent for using tracking technologies and collecting personal data. As a result, webpages appear very different once users provide their consent. This has vast implications on the understanding of Web tracking, on webpage characteristics, on performance measurement, and any other measurement based on Web crawling.

In this paper, we engineered *Priv-Accept*, a tool that automatically crawls websites accepting the privacy policy when a Privacy Banner is found. We run it on a large set of websites popular in Europe and worldwide. Our results highlighted how the picture of the Web varies when measured upon accepting privacy policies: Web Trackers and Third-Parties suddenly become more pervasive, websites more complex and slower to load.

We release *Priv-Accept* as an open-source project. We based it on a set of keywords and, thus, has margins for improvement. We foster its use by the research community to contribute to it and extend our results. We also hope *Priv-Accept* will be included as part of the public projects that provide periodic Web measurements. Our goal is to keep developing *Priv-Accept* to enrich the keyword list, implement additional functionalities, adding the possibility to deny the privacy policies, a much harder task. For this, we envision the design of more sophisticated approaches to manage Privacy Banners, likely based on recent advances in Natural Language Processing and Machine Learning.

ACKNOWLEDGMENTS

The research leading to these results has been funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 871370 (PIMCity project) and the SmartData@PoliTO center for Data Science technologies.

REFERENCES

[1] 2021. AdGuard. https://adguard.com/ (Last accessed May 25, 2021).
[2] 2021. Cliqz AutoConsent. https://github.com/cliqz-oss/autoconsent (Last accessed May 25, 2021).
[3] 2021. Consent-O-Matic. https://github.com/cavi-au/Consent-O-Matic (Last accessed May 25, 2021).
[4] 2021. Docker. https://www.docker.com/ (Last accessed May 25, 2021).
[5] 2021. EasyPrivacy. https://easylist.to/easylist/easyprivacy.txt (Last accessed May 25, 2021).
[6] 2021. HTTPArchive. https://httparchive.org (Last accessed May 25, 2021).
[7] 2021. I don't care about cookies. https://www.i-dont-care-about-cookies.eu/ (Last accessed May 25, 2021).
[8] 2021. Ninja Cookie. https://ninja-cookie.com/ (Last accessed May 25, 2021).
[9] 2021. Remove Cookie Banners. https://chrome.google.com/webstore/detail/remove-cookie-banners/pacehjmodmfilemfbcahnpdcdmlocjnm (Last accessed May 25, 2021).
[10] 2021. SimilarWeb. https://www.similarweb.com (Last accessed May 25, 2021).
[11] 2021. SpeedIndex. https://web.dev/speed-index/ (Last accessed May 25, 2021).
[12] 2021. WhoTracks.me. https://whotracks.me/ (Last accessed May 25, 2021).
[13] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. 2014. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 674–689.

[14] Özgü Alay, Andra Lutu, Miguel Peón-Quirós, Vincenzo Mancuso, Thomas Hirsch, Kristian Evensen, Audun Hansen, Stefan Alfredsson, Jonas Karlsson, Anna Brunstrom, et al. 2017. Experience: An open platform for experimentation with commercial mobile broadband networks. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 70–78.

[15] Waqar Aqeel, Balakrishnan Chandrasekaran, Anja Feldmann, and Bruce M. Maggs. 2020. On Landing and Internal Web Pages: The Strange Case of Jekyll and Hyde in Web Performance Measurement. In *Proceedings of the ACM Internet Measurement Conference (IMC '20)*. Association for Computing Machinery, New York, NY, USA, 680–695.

[16] Alemnew Sheferaw Asrese, Ermias Andargie Walelgne, Vaibhav Bajpai, Andra Lutu, Özgü Alay, and Jörg Ott. 2019. Measuring web quality of experience in cellular networks. In *International Conference on Passive and Active Network Measurement*. Springer, 18–33.

[17] Satya Avasarala. 2014. *Selenium WebDriver practical guide*. Packt Publishing Ltd.

[18] Jan M Bauer, Regitze Bergstrøm, and Rune Foss-Madsen. 2021. Are you sure, you want a cookie?–The effects of choice architecture on users' decisions about sharing private online data. *Computers in Human Behavior* 120 (2021), 106729.

[19] Enrico Bocchi, Luca De Cicco, and Dario Rossi. 2016. Measuring the quality of experience of web users. *ACM SIGCOMM Computer Communication Review* 46, 4 (2016), 8–13.

[20] California State Legislature. 2018. California Consumer Privacy Act of 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375 (Last accessed May 25, 2021).

[21] Council of European Union. 2009. Directive 2009/136/EC amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws. http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32009L0136 (Last accessed May 25, 2021).

[22] Lynne M Coventry, Debora Jeske, John M Blythe, James Turland, and Pam Briggs. 2016. Personality and social framing in privacy decision-making: A study on cookie acceptance. *Frontiers in psychology* 7 (2016), 1341.

[23] Diego Neves da Hora, Alemnew Sheferaw Asrese, Vassilis Christophides, Renata Teixeira, and Dario Rossi. 2018. Narrowing the gap between QoS metrics and Web QoE using Above-the-fold metrics. In *International Conference on Passive and Active Network Measurement*. Springer, 31–43.

[24] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. 2019. Measuring cookies and web privacy in a post-gdpr world. In *International Conference on Passive and Active Network Measurement*. Springer, 258–270.

[25] Hugues de Saxcé, Iuniana Oprescu, and Yiping Chen. 2015. Is HTTP/2 really faster than HTTP/1.1?. In *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 293–299.

[26] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2018. We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. *arXiv preprint arXiv:1808.05096* (2018).

[27] Deloitte. 2020. Cookie Benchmark Study. https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/risk/deloitte-nl-risk-cookie-benchmark-study.pdf (Last accessed May 25, 2021).

[28] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 1388–1401.

[29] Jeffrey Erman, Vijay Gopalakrishnan, Rittwik Jana, and Kadangode K Ramakrishnan. 2015. Towards a SPDY'ier mobile web? *IEEE/ACM Transactions on Networking* 23, 6 (2015), 2010–2023.

[30] European Parliament and Council of European Union. 2016. Directive 95/46/EC. General Data Protection Regulation. http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf (Last accessed May 25, 2021).

[31] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. 2014. The rise of panopticons: Examining region-specific third-party web tracking. In *International Workshop on Traffic Monitoring and Analysis*. Springer, 104–114.

[32] Jens Grossklags and Nathan Good. 2007. Empirical studies on software notices to inform policy makers and usability designers. In *International Conference on Financial Cryptography and Data Security*. Springer, 341–355.

[33] Philip Hausner and Michael Gertz. 2021. Dark Patterns in the Interaction with Cookie Banners. *arXiv preprint arXiv:2103.14956* (2021).

[34] Maximilian Hils, Daniel W. Woods, and Rainer Böhme. 2020. Measuring the Emergence of Consent Management on the Web. In *Proceedings of the ACM Internet Measurement Conference (IMC '20)*. Association for Computing Machinery, New York, NY, USA, 317–332.

[35] Xuehui Hu and Nishanth Sastry. 2019. Characterising third party cookie usage in the EU after GDPR. In *Proceedings of the 10th ACM Conference on Web Science*. 137–141.

[36] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. 2018. Tracing cross border web tracking. In *Proceedings of the Internet Measurement Conference 2018*. 329–342.

[37] Célestin Matte, Nataliia Bielova, and Cristiana Santos. 2020. Do Cookie Banners Respect my Choice?: Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 791–809.

[38] Johan Mazel, Richard Garnier, and Kensuke Fukuda. 2019. A comparison of web privacy protection techniques. *Computer Communications* 144 (2019), 162–174.

[39] Hassan Metwalley, Stefano Traverso, Marco Mellia, Stanislav Miskovic, and Mario Baldi. 2015. The online tracking horde: a view from passive measurements. In *International Workshop on Traffic Monitoring and Analysis*. Springer, 111–125.

[40] Ravi Netravali, Anirudh Sivaraman, Somak Das, Ameesh Goyal, Keith Winstein, James Mickens, and Hari Balakrishnan. 2015. Mahimahi: Accurate Record-and-Replay for HTTP. In *2015 USENIX Annual Technical Conference (USENIX ATC 15)*. USENIX Association, Santa Clara, CA, 417–429.

[41] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P Markatos. 2021. User Tracking in the Post-cookie Era: How Websites Bypass GDPR Consent to Track Users. *arXiv preprint arXiv:2102.08779* (2021).

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

[43] Enric Pujol, Oliver Hohlfeld, and Anja Feldmann. 2015. Annoyed users: Ads and ad-block usage in the wild. In *Proceedings of the 2015 Internet Measurement Conference*. 93–106.

[44] Valentino Rizzo, Stefano Traverso, and Marco Mellia. 2021. Unveiling web fingerprinting in the wild via code mining and machine learning. *Proceedings on Privacy Enhancing Technologies* 2021, 1 (2021), 43–63.

[45] Vaspol Ruamviboonsuk, Ravi Netravali, Muhammed Uluyol, and Harsha V Madhyastha. 2017. Vroom: Accelerating the mobile web with server-aided dependency resolution. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*. 390–403.

[46] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. 2019. Can i opt out yet? GDPR and the global illusion of cookie control. In *Proceedings of the 2019 ACM Asia conference on computer and communications security*. 340–351.

[47] Ashiwan Sivakumar, Shankaranarayanan Puzhavakath Narayanan, Vijay Gopalakrishnan, Seungjoon Lee, Sanjay Rao, and Subhabrata Sen. 2014. Parcel: Proxy assisted browsing in cellular networks for energy and latency reduction. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*. 325–336.

[48] Jannick Sørensen and Sokol Kosta. 2019. Before and after gdpr: The changes in third party presence at public and private european websites. In *The World Wide Web Conference*. 1590–1600.

[49] Stefano Traverso, Martino Trevisan, Leonardo Giannantoni, Marco Mellia, and Hassan Metwalley. 2017. Benchmark and comparison of tracker-blockers: Should you trust them?. In *2017 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 1–9.

[50] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 2019. 4 years of EU cookie law: Results and lessons learned. *Proceedings on Privacy Enhancing Technologies* 2019, 2 (2019), 126–145.

[51] Phani Vadrevu and Roberto Perdisci. 2019. What You See is NOT What You Get: Discovering and Tracking Social Engineering Attack Campaigns. In *Proceedings of the Internet Measurement Conference (IMC '19)*. Association for Computing Machinery, New York, NY, USA, 308–321.

[52] Pelayo Vallina, Álvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. 2019. Tales from the Porn: A Comprehensive Privacy Analysis of the Web Porn Ecosystem. In *Proceedings of the Internet Measurement Conference (IMC '19)*. Association for Computing Machinery, New York, NY, USA, 245–258.

[53] Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Xavier Blanc. 2020. FP-Crawlers: studying the resilience of browser fingerprinting to block crawlers. In *MADWeb'20-NDSS Workshop on Measurements, Attacks, and Defenses for the Web*.

[54] Tony Vila, Rachel Greenstadt, and David Molnar. 2003. Why we can't be bothered to read privacy policies models of privacy economics as a lemons market. In *Proceedings of the 5th international conference on Electronic commerce*. 403–407.

[55] Xiao Sophia Wang, Aruna Balasubramanian, Arvind Krishnamurthy, and David Wetherall. 2014. How Speedy is SPDY?. In *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*. USENIX Association, Seattle, WA, 387–399.

[56] Xiao Sophia Wang, Arvind Krishnamurthy, and David Wetherall. 2016. Speeding up Web Page Loads with Shandian. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*. USENIX Association, Santa Clara, CA, 109–122.