The Internet with Privacy Policies: Measuring The Web Upon Consent

The Internet with Privacy Policies: Measuring The Web Upon Consent

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

To protect user privacy, legislators have regulated the use of tracking technologies, mandating the acquisition of users' consent before collecting data. As a result, websites started showing more and more consent management modules – i.e., Consent Banners – the visitors have to interact with to access the website content. Since these banners change the content the browser loads, they challenge web measurement collection, primarily to monitor the extent of tracking technologies, but also to measure web performance. If not correctly handled, Consent Banners prevent crawlers from observing the actual content of the websites.

In this paper, we present a comprehensive 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 Consent Banners, as most users would do in practice. It lets us compare how webpages change before and after accepting such policies, if present. Our results show that all measurements performed ignoring the Consent Banners offer a biased and partial view of the Web. After accepting the privacy policies, web tracking is far more pervasive, webpages are larger and slower to load.

CCS Concepts: • **Networks** → **Network measurement**; **Network measurement**; • **Security and privacy** → **Human and societal aspects of security and privacy**;

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

ACM Reference Format:
Nikhil Jha, Martino Trevisan, Luca Vassio, and Marco Mellia. 2022. The Internet with Privacy Policies: Measuring The Web Upon Consent. 1, 1 (August 2022), 26 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 personalized content [14, 45, 48]. Hundreds of tracking platforms exist, with many of them gathering information from a large base of users and websites [33, 38, 43, 47].

This picture has created tension over online privacy, and regulatory bodies have started governing the scenario. Lastly, in May 2018, the EU introduced the General Data Protection Regulation (GDPR) [32]. 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) [21] 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

Authors' addresses: Nikhil Jha, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy, nikhil.jha@polito.it; Martino Trevisan, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy, martino.trevisan@polito.it; Luca Vassio, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy, luca.vassio@polito.it; Marco Mellia, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy, marco.mellia@polito.it.

module – the commonly called Consent 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 [14, 30, 33, 37, 38, 41, 43, 45, 47, 47, 48, 53, 55] and performance [15, 17, 20, 27, 31, 44, 49, 51, 59]. These measurements are typically carried out with headless browsers that access webpages and automatize the collection of metadata and statistics. However, today, these measurements could be biased and unrealistic, with the crawler observing possibly very different content than what a user would get after accepting the privacy policies. In fact, the Consent Banners may hide the actual page content, and the browser may load additional content only after the privacy policy acceptance. While researchers have shown the importance of carefully choosing which webpages to test [16], to the best of our knowledge, we are the first to consider the impact of Consent Banners on automatic measurements.

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 exposed and obtain thus realistic figures on performance. The non-standard way of displaying the Consent 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 Consent Banner in about 90% of cases when present.

Armed with *Priv-Accept*, we run an extensive measurement campaign. We focus primarily 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, activating trackers and personalizing ads only after consent is collected. A researcher collecting statistics by crawling the Web without managing consent could erroneously think that tracking is decreasing with respect to the past [37]. 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 Web crawlers. Similarly, after accepting privacy policies, webpages become more complex and heavier since the browser has to load more objects from more third-party servers. Thus, they are slower to load, so that webpages embedding many trackers and ads double or triple the page load time.

Recently, authors of [16] showed how important it is to extend the crawling to internal pages. Here, we show that it is also fundamental to correctly handle the Consent Banners when running extensive Web measurements. For this, we offer *Priv-Accept* as an open-source tool to incentivize other researchers to contribute. Similarly, we offer all the data we collected for this study and the code to generate the figures to the community in an effort to support reproducibility and encourage other studies.[1]

After discussing the scenario and related work in Section 2, we present and test *Priv-Accept* 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 and limitation in Section 7, we summarize our findings in Section 8.

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

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

2 BACKGROUND AND RELATED WORK

Content providers on the Web often monetize the content they offer by 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 web users [43, 47]. Trackers exploit cookies and advanced techniques to enable the collection of personal information [14, 45, 48].

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. In 2013, the European Cookie Law [22] entered into force, which mandates websites to ask for informed consent before using any profiling technology. Later, in May 2018, the General Data Protection Regulation (GDPR) [32] 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 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 Consent Banner implementation [40] and a large fraction have been shown to use tracking technologies before user consent [50, 54]. In the US, the California Consumer Privacy Act (CCPA) [21] 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 Consent Banners [28] and many adopt Consent Management toolsets [36], 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 loaded and visible.

2.2 The Effect of Consent Banners on Web Measurements

Despite cases of misuse, the new regulations had a large impact on the web users and complicate 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 tracker and no ad being loaded. Hu *et al.* [37] 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

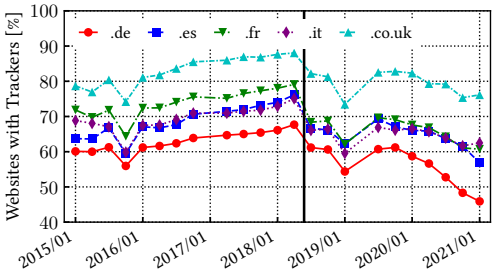


Fig. 2. Percentage of websites containing at least one tracker for five European Top-Level domains (from HTTPArchive). The black vertical line indicates the entry into force of the GDPR. Since then, the apparent pervasiveness of tracking decreased.

measure no significant reduction in long-term numbers of third-party cookies. Dabrowski *et al.* [25] draw similar conclusions, finding an apparent decrease in the use of persistent cookies from 2016 to 2018. Sorensen *et al.* [52] 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).[2] We restrict the analysis on those websites that exist for the whole six years-long periods (9 196 website in total).

Figure 2 could suggest 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 third-party trackers and 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 [14, 16, 30, 33, 37, 38, 43, 45, 47, 48, 55] and counter-measurements [41, 47, 53]. Moreover, this will also impact those works that rely on crawlers and headless browsers [18] to quantify the impact in the wild of new technologies like SPDY, HTTP/2 [20, 27, 31, 59], 4G/5G [15, 17], accelerating proxies [49, 51, 60], or generic benchmark solutions [44]. At last, even spiders and mirroring tools like Wayback Machine and 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

Vallina *et al.* [56] are the first to consider the impact of the Consent Banner presence. First, they instruct a custom OpenWPM crawler to identify specific Consent 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 generalize.

[2]The Top-Level Domain can sometimes be an inaccurate proxy for a website's country. Here, our goal is only to provide a qualitative picture.

Recently, authors of [16] 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 complications due to Consent Banners. Here, we aim at providing an extensive and thorough study of their impact on the Web. Our goal is to enable the automatic study of webpage characteristics as visitors would experience, assuming that most of them accept the default privacy setting as offered by the Consent Banner. Indeed, it has been shown that most users tend to ignore privacy-related notices [23, 34, 58]. Considering GDPR Consent Banners, users tend to accept privacy policies when offered a default button via intrusive banners that nudge users [19, 29], which is often the case [35] with websites presenting large pop-ups or wall-style banners that cover most of the webpage as seen in Figure 1.

For completeness, notice that cookies are among the simplest tracking mechanisms. Authors of [45] show how practices like cookie synchronization, cookie leaking, and other profiling techniques like canvas fingerprinting are common in today's Web. Similarly, authors of [39] show how the crawling context, in terms of vantage point and browser configuration, has a significant impact on the results. Our work is orthogonal to these to obtain automatic, realistic, reliable and user-centric measurements of the Web.

Focusing on automatic management of Consent Banners, some browser add-ons try to hide them by using a list of CSS selectors of known Consent 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 Consent 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*. However, they are based on a list of actions the browser automatically runs when finding a set of popular Consent Management Platforms (CMPs), limiting their effectiveness. In Section 3.2, we compare *Priv-Accept* with Consent-O-Matic – the most mature tool – showing that *Priv-Accept* offers a much higher coverage. Indeed, the diversity of the Web ecosystem, the presence of multiple languages and the fully customizable choice of Consent 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 Consent Banner and automatically accept privacy policies. We aim at a practical and effective approach to accept privacy policies through the offered button.

To illustrate *Priv-Accept* operation, consider again Figure 1. A large Consent Banner appears on the first 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 the background. We refer to these two types of visits as *Before-Accept* and *After-Accept* in the remainder of the paper.

We implement *Priv-Accept* using the Selenium browser automation tool [18], the de-facto standard for browser automation, using Google Chrome as browser. Given a target URL, *Priv-Accept* carries out the following tasks:

(1) It navigates 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-Accept* to the website.
(2) It inspects the Document Object Model (DOM) of the rendered webpage to find a possible *Accept-button* in a Consent Banner. For this, it matches a list of keywords on the text of each node of the DOM. We identify an *Accept-button* if we exactly match any of these keywords. For robustness, we remove leading/trailing/repeated blank characters and the match is performed ignoring the case. We do not use stemming, lemmatization or other techniques for text processing given the specificity of the words to match and the need to support multiple languages.
(3) If *Priv-Accept* finds the *Accept-button*, it clicks on the corresponding DOM element (typically a `<button>`, `<href>` or `<span>` element) to accept the privacy policy and logs the success acceptance.

In the beginning, we built *Priv-Accept* to look for accept buttons through CSS selectors combined with keywords as done in [56] 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 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 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.

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 can be configured to run a:

- *Warm-up visit*: to populate the browser cache.
- *Before-Accept*: to collect statistics on the webpage before accepting the privacy policy, as a Naive Crawler would do.
- *After-Accept*: 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.[3] We visit internal pages both if *Priv-Accept* finds the *Accept-button* and if it does not.

For each page visit, *Priv-Accept* collect several metadata. Considering QoE metrics, here we focus on the Page Load Time, or *OnLoad* time [24]. It allows us to compare the webpage rendering performance with and without privacy policy acceptance. It is simpler and faster to compute than the SpeedIndex [11], allowing large scale measurements. Notice that we neglect metrics that are not affected by the presence of a Consent Banner, such as the Time-to-first-byte (TTFB).

Notice that the *After-Accept* visit can only occur with a warm browser cache in real cases since the browser would have first to complete the *Before-Accept* visit. To fairly compare a *Before-Accept* and *After-Accept*, in our experiments we run a preliminary *Warm-up visit* before the *Before-Accept* to fill the browser cache. This lets us appreciate the eventual extra time to load additional components and fairly compare the *OnLoad* on the two visits with the hot cache. Alternatively, *Priv-Accept* can erase the HTTP cache and clean the socket pool upon each visit to compare webpage performance with a cold cache.

[3]We define internal links as those having the same Fully Qualified Domain Name as the visited website.

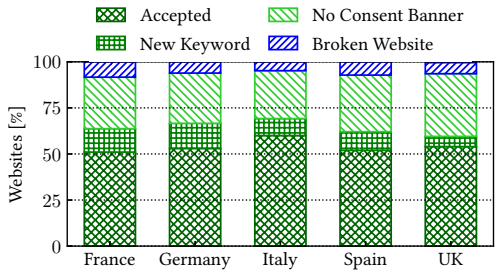


Fig. 3. Validation results of *Priv-Accept* over 200 randomly picked websites per country. Upon two rounds of keyword selection, *Priv-Accept* 92%-95% accurate.

Priv-Accept follows possible redirects during the visits and cases when a script triggers a reload of the webpage. This allows us to manage cases in which the consent banner is hosted on a separate specific landing page than the actual website home page. 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 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`.[4]

3.1 Keyword Selection and Validation

At the core of *Priv-Accept* there 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 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[5], plus the US – which we use as an example of a large, extra-EU country were privacy laws are in force. 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 Consent Banner. We randomly choose half of these websites and manually visit them (from Europe) to extract the accept keyword. In total, we visit 500 websites and identify 186 unique keywords. We next instruct *Priv-Accept* to visit the other half of websites and let it accept the privacy policy, if found. For those where it fails (233 cases), we manually visit them to check i) if they have a Consent Banner, and ii) eventually to extract new keywords. With this, we identify 36 new keywords, 222 in total. During these steps, we also check that the tool correctly accepts the policy.

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, 1000 websites in total. We let *Priv-Accept* visit them and manually check the subset of 448 websites for which *Priv-Accept* did not find (and accepted) a privacy policy. We depict the results in

[4]The containerized version is available on Docker Hub as *martino90/priv-accept*.
[5]In January 2021 UK has enforced the UK GDPR, with practically identical requirements.

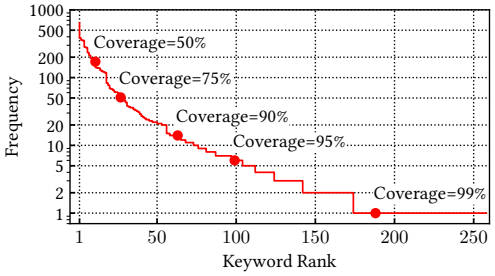


Fig. 4. Frequency of the *Priv-Accept* keywords, with indication of the coverage at different points. The top-98 keywords already cover 95% of websites.

Figure 3. *Priv-Accept* can accept the privacy policy in more than half of websites, independently from the language. In 6 – 14% of cases, we find 36 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 Consent Banner. At last, *Priv-Accept* fails to accept privacy in only 5 – 8% of cases. Investigating further, this is due to some non-standard behavior 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 [57]. For completeness, cases of *False Positives* – i.e., *Priv-Accept* clicking on a wrong DOM element – are possible, although we have not observed any in our manual validation tests.

At the end of the keyword list building phases, we collect a total of 258(186 + 36 + 36) keywords obtained by manually visiting 1181(500 + 233 + 448) websites, covering 6 languages.[6] In Figure 4, we show the distribution of keyword appearance frequency across the entire set of 12 277 Similarweb websites (see Section 3.3 for details on this list). The most common keyword is the string "Ok". Red dots indicate the portion of websites covered by the top-N keywords – i.e., the coverage of the top-N words. The top keywords are very common (note the logarithmic scale on the y-axis), with the top-10 that cover half of the websites. The top-98 keywords cover 95% of the websites, while the remaining appear less than 10 times each in the whole website set. Clearly, we expect the list of keywords to naturally grow as the tail of the Figure 4 suggests. Notice indeed that more than 80 keywords have been found on a single website. Curiously, we find complex strings like "I'm fine with this" or "Alle auswählen, weiterlesen und unsere arbeit unterstützen".[7]

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 Platforms (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 Consent Banners for 35 CMPs. To gauge its performance, we visit the top-100 most popular websites with a Consent 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 Consent 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.

[6] In Spain, some websites are in Catalan, rather than in Spanish.
[7] Which translates to "Select all, keep reading and support our work".

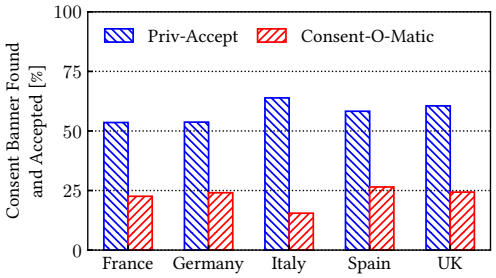


Fig. 5. Privacy policy acceptance rate of *Priv-Accept* and Consent-O-Matic on 100 websites per country. *Priv-Accept* can find and accept Consent Banners on twice as many websites as Consent-O-Matic.

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* accepts 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 Consent Banner, fail to load, or use an unknown keyword. This testifies that the customization of Consent 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 check the impact of using *Priv-Accept* when doing large web measurement experiments. We targets a large set of websites popular in France, Germany, Italy, Spain and US, using a test server located in our university campus. For each of the 6 countries, we use the Similarweb lists to select the top-100 websites from 24 different categories – see Figure 10. These are the top-level unique categories listed in the Similarweb page [13]. In total, we include 12 277 unique websites to visit (as the lists in different countries partially overlap). When visiting websites of a given country, we set the `Accept-Language` header to indicate the appropriate locale and country language. This behavior can be configured in the *Priv-Accept* configuration to allow further experimentation.

We run *Priv-Accept* on 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-Accept* and *After-Accept* 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 [16]. 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. To understand whether Consent Banners appear or have a different impact depending on the visitor location, 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. Here, we aim to check if websites behave differently based on the location of the visitors. Since we are using cloud servers, targeted websites may wrongly recognise the test machines as not regular users and located them in a generic or wrong country. While we cannot check this, we verified that the two most popular commercial

IP location databases (IP2Location[8] and MaxMind[9]) map the IP addresses of our crawlers to the correct country.

To offer a view on a larger number of websites, we visit the top-100 000 websites according to the Tranco list [46]. Unfortunately, the Tranco list does not offer a per-category and per-country rank. We run two separate test sequences: with warm caches, doing (i) *Warm-up visit*, (ii) *Before-Accept*, and (iii) *After-Accept*. And with cold caches, (i) *Before-Accept*, (ii) erase HTTP cache and clean socket pool and (iii) *After-Accept*. Following this procedure, we ensure a fair comparison between *Before-Accept* and *After-Accept* in the two scenarios. Recall that *Priv-Accept* allows one to generate any combination of test sequence with warm/cold cache.

To observe how the presence of trackers changes, 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] (a popular ad-blocking tool). For robustness, we merge the three lists and consider as a potential tracker any third-party domain that appear in at least two lists. In total, we obtain 1 497 domains that we consider tracking services.[10] We finally 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 [54] – 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.

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-Accept* or *After-Accept*. *Priv-Accept* found and accepted a Consent Banner on 63.2% of websites. Here, we aim at quantifying the impact of privacy policy acceptance on European websites (10 542 in total) and 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.[11] With little surprise, the most pervasive Third-Party is `google-analytics.com`. It grows from 61% to 74% in popularity on the *After-Accept*. This value is surprisingly similar to what Metwalley *et al.* [42] found in 2016, when they found `google-analytics.com` appearing in 71% of websites. 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-Accept*, likely being not included in the mechanisms of Consent 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

[8]https://www.ip2location.com/

[9]https://www.maxmind.com/

[10]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`.

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

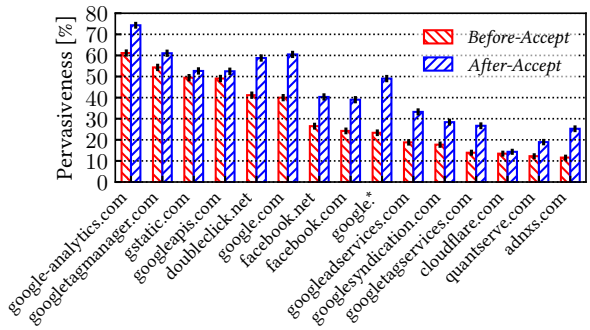


Fig. 6. Pervasiveness of the top-15 Third-Parties (percentage of sites they are in) on 10 542 websites popular in Europe. Most of them are far more pervasive on the *After-Accept*. 95% confidence intervals are reported on each bar.

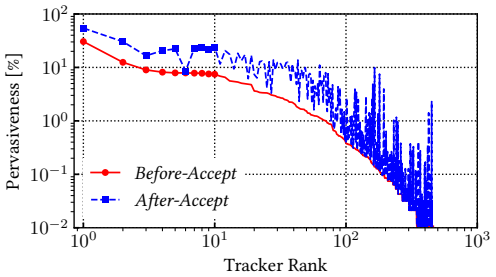


Fig. 7. Pervasiveness of the 342 identified Trackers (percentage of sites they are in) in 10 542 websites popular in Europe. Note that the figure has log-log axes to better show the large variability of Tracker popularity. Also unpopular Trackers result more pervasive on the *After-Accept*.

most popular ones, with pervasiveness growing from 41% to 58% and from 24% to 39% on the *After-Accept*, respectively. They are present in more than twice the number of websites than their first competitor (`quantserve.com`). In Figure 6, we also report 95% confidence intervals. It results that the sample proportion (in percentage) of pervasiveness of Third-Parties is an unbiased estimator of the probability p of a Bernoulli random variable. Therefore, by repeating a number of occurrences of a Bernoulli random variable equal to the number of samples, we obtain the number of successes of a binomial random variable. The confidence intervals become the classical binomial proportion confidence intervals. For the sake of completeness, we report error bars also in the following plots. Note, that, given the large number of samples, the confidence intervals are very narrow and not overlapping between *Before-Accept* and *After-Accept*, except for the case of `cloudflare.com`.

Focusing now on Trackers only, we show their pervasiveness in Figure 7. We count 342 of them. The red curve shows the pervasiveness on the *Before-Accept*, which is what a naive crawler would report. The blue curve shows how the figure changes on the *After-Accept*. The Trackers on the x-axis are sorted in descending order according to their pervasiveness on the *Before-Accept*– hence the *Before-Accept* curve is monotonically decreasing, while the *After-Accept* is not. The increase in pervasiveness is general and includes both popular and infrequent Trackers, reaching one order of

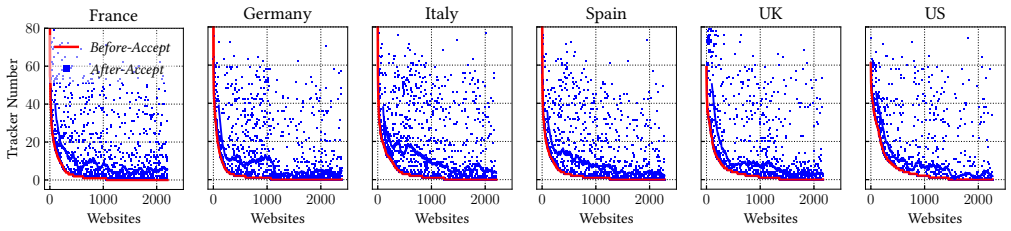


Fig. 8. Trackers per website seen on the landing page. Websites (top 2 500 per country) are sorted by Tracker number on the *Before-Accept* (red curve). The blue points report the number of Trackers in the *After-Accept* for the same websites considered in the red curve, while the blue line represent a moving average with a 100-website window.

magnitude in a some cases. On the *After-Accept*, the number of Trackers that are present on 1% or more of websites grows from 40 to 90. Here, 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 pervasiveness increases.

As it emerges from Figure 7, many Trackers are widespread even on the *Before-Accept*. This hints at a possibly wrong implementation of the GPDR regulation, which mandates acquiring the visitor's explicit consent before activating any tracking mechanisms. To be precise, the presence of Trackers on the *Before-Accept* does not necessarily entail a violation of the law. An analysis of the most popular cookies reveals the presence of test cookies during the *Before-Accept* using a form similar to `test_cookie = CheckForPermission`. Google Analytics is a notable example. 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-Accept* pervasiveness of some Trackers includes cases in which only test cookies are actually used (curiously with expiration date longer than a month). Here we limit to observe that often Trackers set some (potentially) profiling cookies even on the *Before-Accept*.

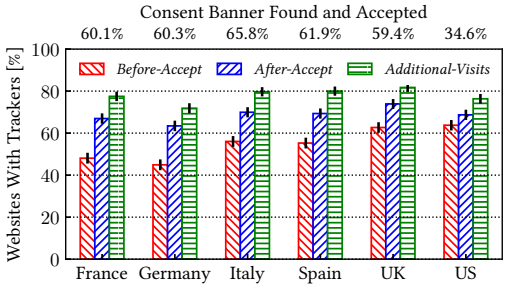
Take away: *Collecting measurements with or without consent to privacy policies leads to a largely different picture. Upon consent, Trackers are far more pervasive than it appears beforehand.* Priv-Accept *is instrumental for this goal, thanks to its ability to handle Consent Banners and accept website privacy policies.*

4.2 Breakdown on Websites

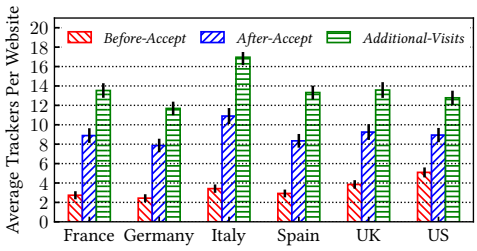
We now detail the impact of accepting privacy policies on the number of Trackers found in each website, breaking down our results by country and website category.

4.2.1 Analysis by country. Figure 8 shows websites sorted in descending order by the number of contacted Trackers as measured in the *Before-Accept* (red curve). This number tends to grow on the *After-Accept* (blue points), where we observe some websites that present 50-70 more Trackers. To increase readability, in Figure 8, the blue line reports the moving average (with a 100 window) of the number of contacted Trackers on the *After-Accept*. Curiously, some websites that already include Trackers in the *Before-Accept* include more Trackers in the *After-Accept*. This again may hint at a wrong implementation of the Consent Banner, which fails to hinder the presence of offending Trackers. The increase is less remarkable for US-popular websites – mainly due to the less widespread presence of Consent Banners.

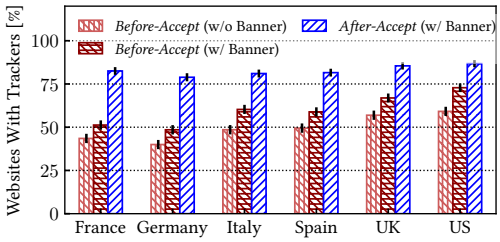
To better quantify Tracker presence, we show the fraction of websites containing at least one Tracker in Figure 9a. As in Figure 6, we report 95% confidence interval on these sample proportions.



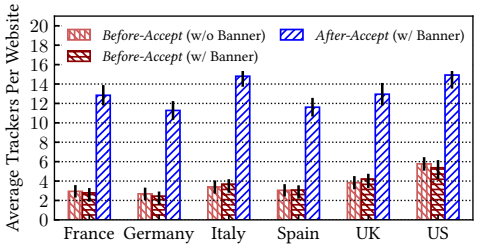
(a) Percentage of websites embedding Trackers. The top x-axis details the fraction of websites in such category where *Priv-Accept* found and accepted privacy policies.



(b) Average number of Trackers per website.



(c) Percentage of websites embedding Trackers, splitting websites with and without a Consent Banner.



(d) Average number of Trackers per website, splitting websites with and without a Consent Banner.

Fig. 9. Tracker penetration during different phases of a browsing sessions (top 2 500 websites per country). 95% confidence intervals are reported on each bar. On the *After-Accept* and *Additional-Visits*, we find many more Trackers.

About 50% of websites popular in European countries already include at least one Tracker on *Before-Accept*. This happens more frequently in the UK (63%) and less often in Germany (44%). Again, note that a website embedding a Tracker on the *Before-Accept* does not necessarily represent a violation of the GDPR, even if this can often be the case [54]. Interestingly, in the US this figure is higher than in European countries. Recalling that the probability of encountering a Consent Banner in the US is lower, this hints at a positive effect of the GDPR on popular European websites. The percentage of websites containing Trackers in the *After-Accept* grows for all European countries from a +11% increase in the UK to +20% for Germany. Confidence intervals never overlap. This increase is moderate (+5%) in the US, given the lower fraction of those websites having a Consent Banner. We complete this analysis by reporting how this fraction increases when performing 5 *Additional-Visits* as recommended in [16]. Our results confirm this need, with the chance to observe at least one Tracker that further grows by 5%-10% in *Additional-Visits* when compared to the *After-Accept*. Note that, considering each country, none of the confidence intervals overlap between *Before-Accept* and *After-Accept* and between *After-Accept* and *Additional-Visits*.

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. Also in this case we report 95% confidence intervals. The sample mean is an unbiased estimator of the true mean, and we can derive confidence intervals through central limit theorem. For all

countries, the average amount of Trackers more than doubles on the *After-Accept*, and performing *Additional-Visits* further increases this figure (with non-overlapping confidence intervals). In Italy, for instance, this figure grows by a factor of 4 when comparing *Before-Accept* 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. This hints that popular websites in the United States may be 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. Afterwards, 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.

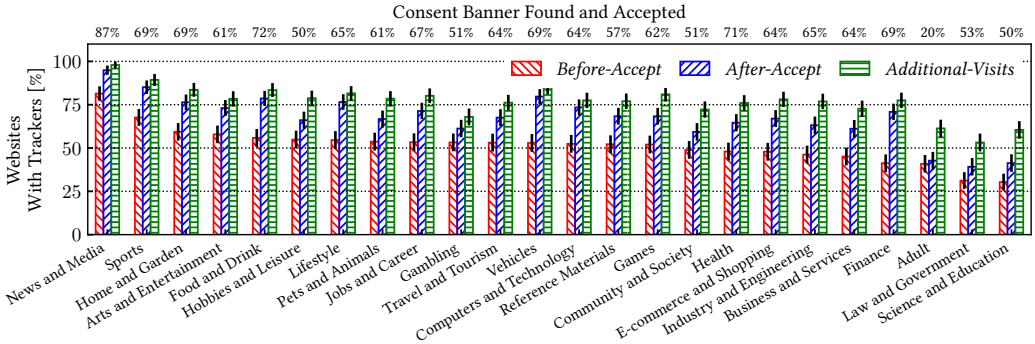
To appreciate the variation in the number of Trackers for those websites implementing a Consent Banner, we deepen the analysis by showing separately websites for which *Priv-Accept* has found (or not) a Consent Banner. Our goal is to show how Tracker number varies on the *Before-Accept* and *After-Accept* for those websites implementing the Consent Banner. Figure 9c shows the percentage of websites with at least one Tracker, and Figure 9d shows the number of Trackers per website. The dark red bars and blue bars show results on the *Before-Accept* and *After-Accept* for those websites where *Priv-Accept found* a Consent Banner. As before, the increase of Trackers is sizeable. For completeness, the light red bars report the same measure for those websites where *Priv-Accept did not* find any Consent Banner.

We finally observe that the probabilistic nature of Web tracking and bidding mechanisms results in a different number of Trackers contacted at each visit. To obtain the most reliable measurements, we test each website 5 times, each time visiting 5 internal pages. We note 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-Accept* per website, overall, we find 69.1% of them containing one (or more) Trackers. Repeating 5 times the test and considering whether we find at least one Tracker among all visits, this percentage increases only to 70.0%. Similarly, the average number of Trackers (as in Figure 9b), increases from 6.5 to 7.8. We report additional details on this in the Appendix and in Figure 15.

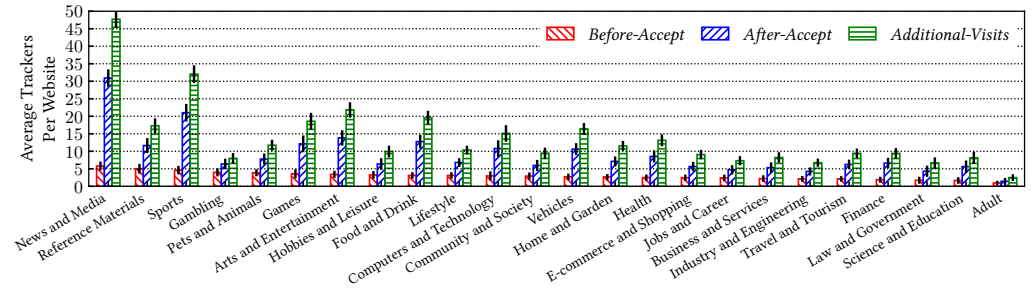
4.2.2 Analysis by category. We now break down the picture by category, showing the results in Figure 10. We explicitly target websites of 24 categories, each containing the top-100 websites for the considered countries.

Starting from Figure 10a, we report the percentage of websites of a given category that contain at least one Tracker. As before, there is a large increase from *Before-Accept* to *After-Accept*. Exceptions are the *Adult*, *Law and Government* and *Gambling* categories, where the confidence intervals overlap. For *Adult* this is likely due to the low number of websites with Consent Banners (20%) and confirms the peculiarity of the tracking ecosystem on Adult websites [56]. As previously observed in Figure 9a, performing *Additional-Visits* further increases the chance of encountering at least one Tracker, even though in this case the increase is limited and we observe some overlaps between *After-Accept* and *Additional-Visits* confidence intervals.

Moving to the number of trackers per website shown in Figure 10b, we observe large increase in the *After-Accept* case, confirming that most Trackers appear only after the user accepts the privacy policies and when visiting internal pages. Here, differences across categories are all pronounced, with those categories that heavily depend on advertisements (*News and Media*, *Sports*, *Games*, *Arts and Entertainment*) that have to rely on a large number of Trackers to support behavioral advertisements. This is noticeable already on the *Before-Accept*. For example, access to a *News* website leads to contact 5.7 Trackers on average in *Before-Accept*. Here, *Priv-Accept* successfully



(a) Percentage of websites embedding Trackers. The top x-axis details the fraction of websites in such category where *Priv-Accept* found and accepted privacy policies.



(b) Average number of Trackers per website.

Fig. 10. Trackers penetration and number on websites (top 2 500 per country) during different phases of a browsing session, separately by category. We sort categories from the highest to the lowest percentage of websites with Trackers in *Before-Accept*. 95% confidence intervals are reported on each bar. In some cases (e.g., News and Media), on the *After-Accept* and *Additional-Visits* the increase is very pronounced.

accepts the privacy policies in 87% of cases. Indeed, being *News* websites very popular, they tend to correctly implement the privacy regulations and to show a well-configured Consent Banner. Upon acceptance, suddenly, the number of Trackers becomes almost 6 times higher (30.9 for *News*) and 9 times higher when doing *Additional-Visits* (47.7 trackers on average). For *Sport*, *Food and Drink* and *Arts and Entertainment* the average number of Trackers more than triples in *After-Accept*. Only for the *Adult* category confidence intervals overlap.

These numbers are particularly interesting if read in the perspective of recent works. Englehardt *et al.* [30], in 2016, measured an average of 35 Trackers per website on News websites. In 2021, we find similar numbers (30.9) on the *After-Accept*, while, due to the spread of Consent Banners, on the *Before-Accept* we would only find 5.7, on average. On Sport category, Englehardt *et al.* [30] measured 27 Trackers per website. In 2021, we find 21.0 on the *After-Accept*, while only 4.6 on the *Before-Accept*. These results well highlight the need for correctly handling the Consent Banners to observe the extensiveness of web tracking. In a nutshell, thanks to *Priv-Accept*, we obtain the fundamentally different figure in the *After-Accept* and *Additional-Visits*.

The case of *Adult* websites is worth a specific comment. *Priv-Accept* finds the Consent Banner on only 20% of them, and a manual check on 50 of them confirms that the large majority of them do not

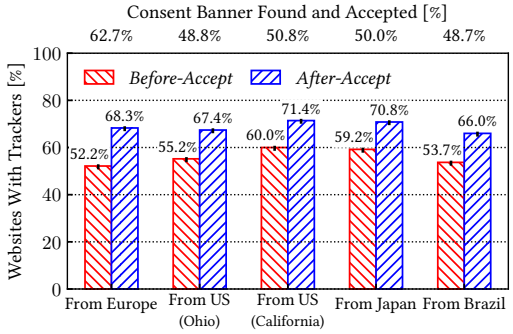


Fig. 11. Websites with Trackers (12 277 from the Similarweb lists) when crawling from different countries. 95% confidence intervals are reported on each bar. From non-European countries, *Priv-Accept* found fewer Consent Banners, but the amount of Trackers on the *After-Accept* is similar. Outside Europe, top-ranked websites tend to include more Trackers.

offer any Consent Banner. Tracking is also limited upon acceptance, and the confidence intervals between *Before-Accept* and *After-Accept* even overlap. Similar results were previously found by Vallina *et al.* [56], where the authors 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 popular tracker lists.

Take away: *Upon consent, the number of Trackers embedded in websites increases by a factor of up to 4 times. European and US websites end up with a similar number of Trackers. The increase is particularly pronounced for certain website categories – for example, News and media or Sport websites – that rely on ads as revenue stream.*

4.3 Visits from Outside Europe

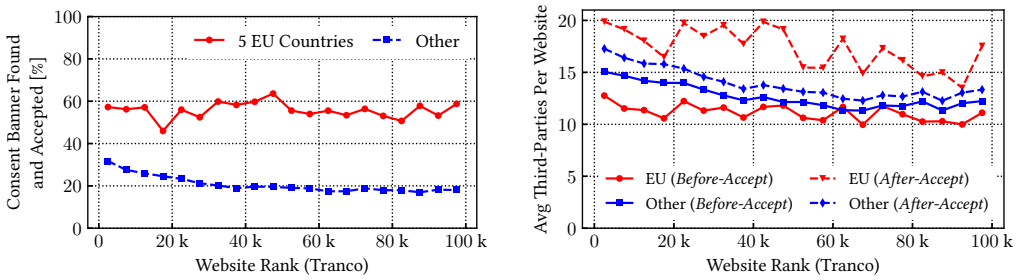
We now consider additional measurement campaigns using crawling servers in the Amazon AWS data centers located in the US (Ohio and California), Japan and Brazil. Figure 11 summarizes our findings. First, notice how *Priv-Accept* accepted privacy policies on around 10% fewer websites (about 1 150 – 1 200) when run from outside Europe, as reported on top x-labels. Checking the screenshot taken by *Priv-Accept* during the visit on a random subset of these websites, we confirm that no Consent Banner is displayed. We can conclude that some websites customize the Consent Banners based on visitors' properties, such as their location. If the visit comes from not EU country, no Consent Banner is shown.

This different behaviour of websites affects also the statistics of the fraction of websites that embed trackers in the *Before-Accept* and *After-Accept* visits. Visiting from outside Europe leads to an increase of Tracking on the *Before-Accept* in all cases, while, on the *After-Accept*, changes are limited.

Take away: *The crawling location location has some impact on the results. This is mostly due to websites that show or not show the Consent Banner based on the user's location, thus not enabling or enabling tracking on the* Before-Accept.

5 IMPACT ON COMPLEXITY AND PERFORMANCE ON TOP-100K WEBSITES

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 may impact the amount of data to download and the rendering performance. Here,



(a) Percentage of websites with a Consent Banner.

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

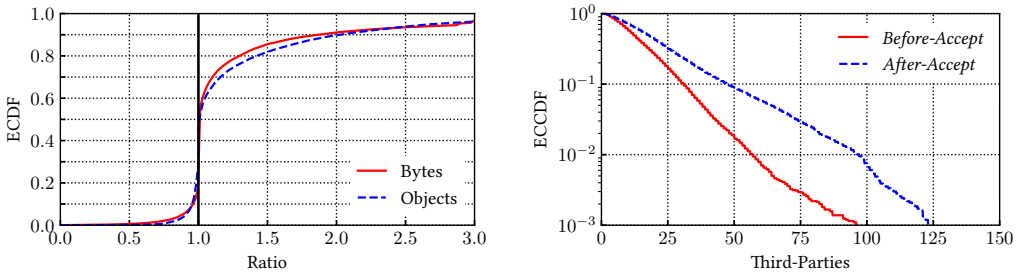
Fig. 12. Percentage of websites with a Consent Banner and average Third-Parties per website over the top-100 k websites in Tranco list, computed every 5 000 websites in the rank. Top websites are more likely to implement a Consent Banner in a *Priv-Accept* supported language.

we do not restrict on a per-country or per-category analysis and use the crawl on the top-100 000 websites according to the Tranco global list.

For each website, we visit only the landing page, doing a *Warm-up* visit to fill the browser cache, followed by a *Before-Accept* and *After-Accept*. 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 worldwide 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 for which regulations do not apply. To give more insights, we detail the percentage of websites with a Consent Banner on the Tranco list in Figure 12a, computed every 5 000 websites in the rank. The solid red line reports the percentage for websites popular in the 5 European countries we target. Websites belong to this set if (i) they appear in the Similarweb ranks for the 5 countries or (ii) the Top-Level Domain belongs to the 5 countries.[12] Out of these 6 917 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%). This percentage does not change with website popularity. Conversely, for the remaining websites (blue dashed line), the share of websites where *Priv-Accept* found a Consent Banner is 32% for the top-ranked and then it settles around 20%, hinting that some globally popular websites tend to implement a Consent Banner even if they are based outside Europe, using a language supported by *Priv-Accept* (likely English). In 2020, Hills *et al.* [36] found that popular CMPs are present on almost 10% of websites in the top-10 k Tranco list. Here, with *Priv-Accept*, we can affirm that Consent Banners (regardless the employed CMP) appear in more than 30% for the same set of websites.

The high number of Consent Banners found for the 5 European countries reflects in a large increase of the number of Third-Parties from the *Before-Accept* to the *After-Accept*, as shown in Figure 12b. The solid red line highlights that these websites already include, on average, 11.1 Third-Parties in the *Before-Accept*. In the *After-Accept*, the average grows to 17.3. Differently, the increase for the non-EU websites is smaller – see the area between the blue solid and dashed lines. In the *Before-Accept*, Third-Parties are larger than for the 5 European countries if we compare the solid blue and red lines. This is due to the larger presence of non-EU websites, which do not have to implement a Consent Banner. In the *After-Accept* (dashed blue line), the increase is moderate, not reaching the values of the 5 European countries (red dashed line), potentially because *Priv-Accept*

[12]The Tranco list does not provide a per-country rank.



(a) Distribution of the page size (in bytes and objects) ratio over all websites.

(b) Distribution of the number of Third Parties. Notice the log scales.

Fig. 13. Webpage characteristic before and upon consent to privacy policies (Tranco list). On the *After-Accept* webpages are larger and include more Third Parties.

misses many *Accept-button* in non-supported languages and of possible custom tracking domains not present in our lists. For the sake of completeness, in the Appendix, we report the same picture as in Figure 12b showing the number of Trackers instead of Third-Parties, providing similar insights.

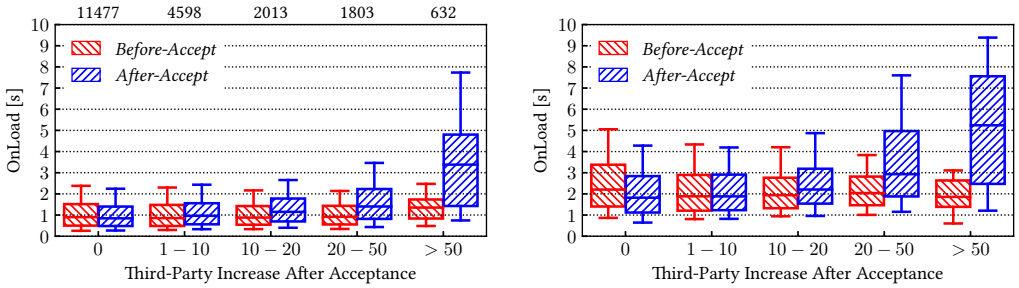
Take away: *For the five European countries considered, the percentage of websites with a Consent Banner (and the number of third parties) is approximately flat with respect to website rank. For the websites of the remaining countries,* Priv-Accept *may miss some* Accept-button *due to the usage of local languages.*

5.1 Impact on Page Objects and Size

We focus on the webpage complexity in terms of bytes and objects to download. We compute the ratio R between the measurement on the *Before-Accept* and *After-Accept*, 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 and objects. As expected, accepting the privacy policy increases the webpage size ($R > 1$) by a sizeable factor. For instance, about 9% of websites download more than twice the objects, and about 5% of websites sees an increase of 3 times or more.

Interestingly, we also observe some websites that are lighter in the *After-Accept* than in the *Before-Accept*. Investigating further, these cases are mostly due to the lack of additional content upon acceptance coupled with the saving of not loading the CMP objects on the *After-Accept*. This happens commonly on those websites that either add a Consent Banner despite not using tracking mechanisms, or that 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-Accept* and *After-Accept*. We show the Empirical Complementary Cumulative Distribution Function (ECCDF) in Figure 13b. On median, websites rely on 12 Third-Parties on the *Before-Accept*. This figure grows to 17 on the *After-Accept*. The ECCDF highlights the tail of the distribution where we observe those websites that rely on a very 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. This growth in the number of Third-Parties is mostly due to an increase of Trackers and objects related to advertisements that gets loaded after accepting the privacy policy. We also perform statistical tests to compare whether the mean and median of the two sample distributions



(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. Website adding many Third Parties on the *After-Accept* are also slower to load.

are statistically different at level 0.05 (t-Test for the mean and Mood's test for the median). Both result statistically significant in *After-Accept*. In Appendix, we include the picture as above, plotting the number of Trackers instead of Third-Parties, leading to similar conclusions.

Take away: *When the Consent Banner is accepted, websites are larger, with* 9% *of them containing more than twice as many objects. Websites including more than* 50 *Third-Parties increase from* 1.8% *to* 9.2%.

5.2 Impact on Page Load Time

The Third-Party domains appearing after acceptance are generally devoted to advertisements, analytics and Web tracking. Contacting them has direct implications on the page load time and, indirectly, on the users' QoE [24]. We thus expect this to cause an increase on the page load time because the browser has to resolve the server name via DNS and contact more servers. For instance, this ultimately limits the advantages offered by new protocols like the stream multiplexing and the header compression offered by HTTP/2 and HTTP/3.

To gauge this, we dissect the webpage load time in Figure 14, comparing separately visits with a warm cache (Figure 14a) and with a cold cache (Figure 14b). In case of warm cache, we run a *Warm-up visit*, then the *Before-Accept* and *After-Accept*. In case of of cold cache, we run the *Before-Accept* without a *Warm-up visit*. Then we erase the HTTP cache and socket pool, then we run the *After-Accept*.

We report the distributions of the *onLoad* time for websites with similar number of additional Third-Parties that are loaded in the *After-Accept*. 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 on the top the respective boxplot. As expected, 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-Accept* and *After-Accept*. Considering visits with cold browser cache (Figure 14a), those 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. Here, 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 seconds, which happens in less than 2% of cases during the *Before-Accept*.

We already observed such an increase in our previous study [53], where we measured that median *onLoad* time increases by 1.3*s* when cookies policies are accepted. We statistically compare all these couples of sample distributions between *Before-Accept* and *After-Accept*, testing differences in the median at a significance level 0.05 (Mood's test). The test is passed in all cases, showing statistically significant differences.

Similar considerations hold for visits with a cold browser cache (Figure 14b). As expected, with the clean cache, websites load time increases – compare values in Figures 14a and 14b. Those that do not add new Third-Parties tend to load slightly faster on the *After-Accept*, potentially due to the absence of the Consent Banner. In fact, differences are statistically significant in the median of the distributions between *Before-Accept* and *After-Accept*, except for the group $1-10$ additional Third-Parties. Again, we observe that those adding several Third-Parties after acceptance have much higher *onLoad* time on the *After-Accept* than on the *Before-Accept*: The median *onLoad* time increases from 1.8 to 5.2 seconds. Finally, we observe that the *onLoad* time values tend to be lower than what measured in older works, potentially because of the advances of content delivery network and increased hardware and software performance. Bocchi *et al.* [20] measured a median *onLoad* time of 3s in 2016 on a similar albeit smaller set of websites.

Take away: *Measuring the webpage load time of websites without considering the implications of accepting the Consent Banners would result in a very biased measurement. Websites that include many more Third-Parties upon acceptance are significantly slower to load.*

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 in 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 – including 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 LIMITATIONS

Our work presents a few limitations, some of which could be addressed in future work.

First, *Priv-Accept* is designed to accept the privacy policy in the Consent Banner. It could be interesting to extend *Priv-Accept* to consider different keywords to choose the different options (e.g., to Opt-Out) on the Consent Banner and verify if websites correctly implement the end-user choice.

Second, the keyword list is manually compiled and static. We leave for future work the design of an automatic mechanism to enlarge and maintain the list. For instance, one can envision a community effort to enrich the list. It would also be interesting to consider some Natural Language Processing-based approaches to compile the keyword list automatically.

Third, currently, *Priv-Accept* uses a global list of keywords, regardless of the website's language. Although unlikely, a keyword may have a different meaning in another language, leading to false positives. A simple solution would be to add support for country- and language-specific lists of keywords.

Considering the results on the web tracking pervasiveness, we here focused on those based on tracking cookies, and we ignore advanced techniques for web tracking such as CNAME cloaking [26], a technique to embed Trackers as first-party domains, or device fingerprinting [48]. Our results

are thus an underestimation of the extensiveness. This problem is general and not specific for *Priv-Accept*.

Moreover, in our experiments, we set the browser language according to the country of each visited website. However, websites may customize their behaviour depending on the users' language, as some are already doing based on the user's location. *Priv-Accept* already allows configuring the content of `Accept-Language` header, making it possible to study this aspect in detail.

8 CONCLUSIONS

In this paper, we demonstrated how the recent regulations had changed the Web landscape, challenging its automatic measurements through traditional Web crawlers. Websites now massively deploy Consent 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 when measuring Web tracking, webpage characteristics, website performance, and any measurement based on Web crawling.

In this paper, we engineered *Priv-Accept*, a tool that automatically crawls websites accepting the privacy policy when a Consent Banner is found. We run it on many websites popular in Europe and worldwide. Our results highlighted how the observed 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, along with the dataset used throughout the paper. 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 more complex task. For this, we envision the design of more sophisticated approaches to manage Consent 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 September 6, 2021).
[2] 2021. Cliqz AutoConsent. https://github.com/cliqz-oss/autoconsent (Last accessed September 6, 2021).
[3] 2021. Consent-O-Matic. https://github.com/cavi-au/Consent-O-Matic (Last accessed September 6, 2021).
[4] 2021. Docker. https://www.docker.com/ (Last accessed September 6, 2021).
[5] 2021. EasyPrivacy. https://easylist.to/easylist/easyprivacy.txt (Last accessed September 6, 2021).
[6] 2021. HTTPArchive. https://httparchive.org (Last accessed September 6, 2021).
[7] 2021. I don't care about cookies. https://www.i-dont-care-about-cookies.eu/ (Last accessed September 6, 2021).
[8] 2021. Ninja Cookie. https://ninja-cookie.com/ (Last accessed September 6, 2021).
[9] 2021. Remove Cookie Banners. https://chrome.google.com/webstore/detail/remove-cookie-banners/pacehjmodmfilemfbcahnpdcdmlocjnm (Last accessed September 6, 2021).
[10] 2021. SimilarWeb. https://www.similarweb.com (Last accessed September 6, 2021).
[11] 2021. SpeedIndex. https://web.dev/speed-index/ (Last accessed September 6, 2021).
[12] 2021. WhoTracks.me. https://whotracks.me/ (Last accessed September 6, 2021).
[13] 2022. SimilarWeb. https://www.similarweb.com/category (Last accessed January 31, 2022).
[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] California State Legislature. 2018. California Consumer Privacy Act of 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375 (Last accessed September 6, 2021).

[22] 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 September 6, 2021).

[23] 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.

[24] 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.

[25] 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.

[26] Ha Dao, Johan Mazel, and Kensuke Fukuda. 2021. CNAME Cloaking-Based Tracking on the Web: Characterization, Detection, and Protection. *IEEE Transactions on Network and Service Management* 18, 3 (2021), 3873–3888. https://doi.org/10.1109/TNSM.2021.3072874

[27] 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.

[28] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. In *26th Annual Network and Distributed System Security Symposium (NDSS '19)*. Internet Society.

[29] Deloitte. 2020. Cookie Benchmark Study. https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/risk/deloitte-nl-risk-cookie-benchmark-study.pdf (Last accessed September 6, 2021).

[30] 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.

[31] 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.

[32] 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 September 6, 2021).

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] Jordan Jueckstock, Shaown Sarker, Peter Snyder, Aidan Beggs, Panagiotis Papadopoulos, Matteo Varvello, Benjamin Livshits, and Alexandros Kapravelos. 2021. *Towards Realistic and ReproducibleWeb Crawl Measurements*. Association for Computing Machinery, New York, NY, USA, 80–91.

[40] 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.

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

[42] Hassan Metwalley, Stefano Traverso, and Marco Mellia. 2016. Using passive measurements to demystify online trackers. *Computer* 49, 3 (2016), 50–55.

[43] 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.

[44] 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.

[45] 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*. Association for Computing Machinery, New York, NY, USA, 2130–2141. https://doi.org/10.1145/3442381.3450056

[46] 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).

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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*.

[58] 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.

[59] 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.

[60] 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.

APPENDIX

Impact of Repeated Visits on Tracking Measurements

We here complement the analysis we carried out on the last paragraph of Section 4.2.1. Web tracking involves a number of mechanisms (real-time bidding among all) that result in the same page containing different Trackers on multiple visits. To obtain a reliable picture, we repeat each test 5 times. In Figure 15, we show how two macroscopic tracking measurements vary with different number of repetited visits for each website. The blue line in the figure shows the fraction of websites that contain at least one Tracker when measured with an increasing number of test repetitions. It is moderately affected by the number of tests, increasing from 69.1% with a single repetition to 70.0% with 5 repetitions. Similarly, the average number of Trackers, increases from 6.5 to 7.8.

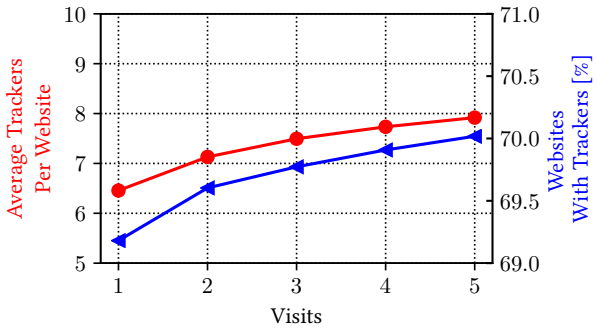


Fig. 15. Variation of tracker number with different numbers of repeated visits. Measurements have sizeable despite moderate variation when repeated.

Trackers per Website (Tranco List)

We here report the same analyses depicted in Figure 12b and Figure 13b showing the number of Trackers instead of the number of Third-Parties. The two pictures lead to similar conclusions.

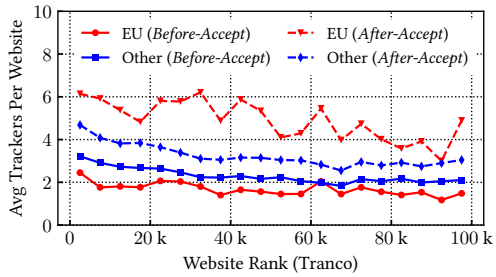


Fig. 16. Average number of Trackers per website (Tranco list).

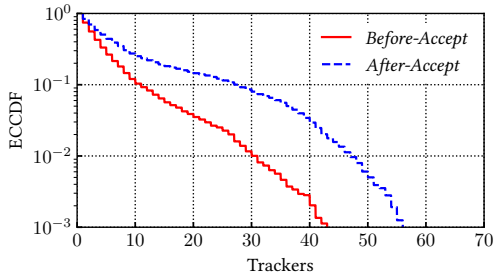


Fig. 17. Distribution of the number of Trackers (Tranco list). Notice the log scales.