Measuring the Emergence of Consent Management on the Web

Maximilian Hils
University of Innsbruck
maximilian.hils@uibk.ac.at

Daniel W. Woods
University of Innsbruck
daniel.woods@uibk.ac.at

Rainer Böhme
University of Innsbruck
rainer.boehme@uibk.ac.at

ABSTRACT

Privacy laws like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have pushed internet firms processing personal data to obtain user consent. Uncertainty around sanctions for non-compliance led many websites to embed a Consent Management Provider (CMP), which collects users' consent and shares it with third-party vendors and other websites. Our paper maps the formation of this ecosystem using longitudinal measurements. Primary and secondary data sources are used to measure each actor within the ecosystem. Using 161 million browser crawls, we estimate that CMP adoption doubled from June 2018 to June 2019 and then doubled again until June 2020. Sampling 4.2 million unique domains, we observe that CMP adoption is most prevalent among moderately popular websites (Tranco top 50-10k) but a long tail exists. Using APIs from the ad-tech industry, we quantify the purposes and lawful bases used to justify processing personal data. A controlled experiment on a public website provides novel insights into how the time-to-complete of two leading CMPs' consent dialogues varies with the preferences expressed, showing how privacy aware users incur a significant time cost.

CCS CONCEPTS

• **Networks** → **Network measurement**; • **Information systems** → **Online advertising**; • **Security and privacy** → *Privacy protections*; Usability in security and privacy.

KEYWORDS

GDPR, CCPA, consent, privacy, web measurement

ACM Reference Format:
Maximilian Hils, Daniel W. Woods, and Rainer Böhme. 2020. Measuring the Emergence of Consent Management on the Web. In *ACM Internet Measurement Conference (IMC '20), October 27–29, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3419394.3423647

1 INTRODUCTION

Vendors harvesting personal data prefer operating beyond the user's attention as evidenced by the use of secret tracking technologies [1, 29, 38]. This was tolerated by websites who rely on advertising revenues [51]. Sanctions associated with recent privacy laws threaten this state of affairs. In the EU, the General Data Protection Regulation (GDPR) requires firms processing personal data to establish a legal basis, such as by obtaining user consent. In the US, the California Consumer Privacy Act (CCPA) requires websites to collect the consent of minors and also to allow users to opt-out of the sale of their personal data. To comply with both laws, an infrastructure of consent must be designed so that users can consent to the privacy practices of websites and Ad-tech vendors.

In the past, each website offered a unique privacy policy and dialogue. This diversity overwhelmed users who could not commit hundreds of hours to reading each privacy policy [6, 36] nor navigate novel interface designs without making errors [2]. Privacy advocates argued that users should set preferences in the browser to avoid such problems [9, 27, 34], whereas Ad-tech companies lobbied against standardized privacy. However, the new imperative to obtain consent creates problems for Ad-tech vendors who must manage and document heterogeneous forms of consent collected across multiple websites.

Consent management providers (CMPs) emerged in the last three years to standardize the collection of online consent. These intermediaries define legal terms and conditions, present these to users via an embedded consent dialogue, store the resulting signal, and share it with third-parties. In essence, CMPs have created a consent ecosystem involving users, websites, and third-party vendors. For example, one CMP allows websites to collect consent for a 'Global Vendor List' with a membership fee of 1200 €, which was termed the commodification of consent [60].

The rise of CMPs represents a new stage in how privacy preferences are communicated, with previous stages including cookies settings in browsers [37] or custom cookie banners on websites [53]. This paper offers a longitudinal study of the formation of a consent ecosystem orchestrated by CMPs. We introduce the notion of a consent flow—from users through consent dialogues to a website and then onto third-parties—and make measurements at each interface. This complements post-GDPR related work relying on snapshots of relatively small samples of domains, which is shown in Figure 1.

Our insights include:

- Using 161 million browser crawls, we measure CMP adoption over time and by website popularity. We show that uptake is most prevalent among 'mid-market' sites ($50^{\text{th}} - 10,000^{\text{th}}$), although this varies between CMPs. We also show the winners and losers of inter-CMP competition in the form of websites switching CMPs.
- In terms of methodology, we introduce a novel URL sampling approach seeded by social media shares, which improves subsite coverage. This is complemented by a traditional toplist sample.

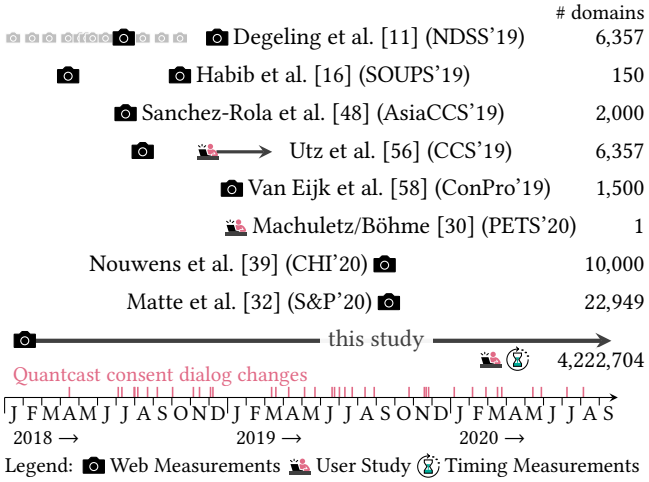

Figure 1: Previous studies conducted point-in-time snapshots of small samples in a rapidly changing environment. For example, the consent prompt of a single CMP (Quantcast) changed 38 times in our observation period.

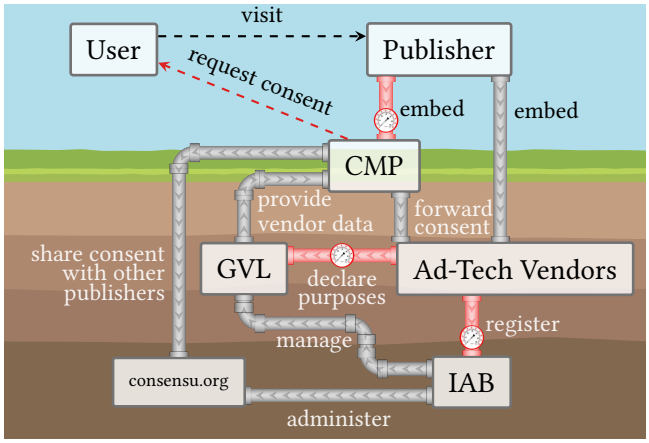

Figure 2: Surfacing the web's new compliance engine: Publishers embed CMPs, which display consent prompts to users, forward consent decisions to ad-tech vendors and also share it globally across websites. In the background, the IAB orchestrates this through its Transparency and Consent Framework (TCF).

- Using APIs from the Ad-tech industry, we quantify the purposes and lawful bases used to justify processing personal data. We find many vendors claiming 'legitimate interest', which allows them to process data without the user's consent.
- We address gaps in the literature by measuring the time to complete consent dialogues, highlighting how users incur a significant time cost when opting out.

Section 2 provides information about the consent ecosystem. Section 3 describes our measurement approach. Section 4 presents our results, which are discussed in Section 5. We identify related work in Section 6 and offers conclusions in Section 7.

2 BACKGROUND

Section 2.1 describes how privacy laws create demand for consent management. Section 2.2 describes the organisations and technical standards relevant to consent management solutions.

2.1 Privacy Laws and Consent

The role of user consent in recent privacy laws is the most significant aspect for this paper. The GDPR applies to all firms processing personal data, which entangles Ad-tech trackers and data brokers as well as websites. Such firms can establish a legal basis for doing so by obtaining user consent (Article 6.1a) or by claiming a legitimate interest (Article 6.1b–f), such as if the data processing protects the "vital interests of the data subject or of another natural person" (6.1d) [47]. If controllers choose to obtain consent, it must be a "freely given, specific, informed and unambiguous indication of the data subject's wishes" (Recital 32) and "documented" (7.1). A data controller infringing either Article 6 or 7 is punishable by "a fine up to € 20 million or up to 4% of the annual worldwide revenue."

In the United States, the California Consumer Privacy Act, which came into effect in January 2020, requires websites to: obtain parental consent for users under 13; affirmative consent for those under 16; and to allow other users to opt-out of the sale of their personal data [17]. The CCPA and GDPR further differ in the obligations on third-party vendors and the definition of personal information. The resulting uncertainty created a business opportunity for CMPs who claim to specialize in compliance. The next section describes the resulting products.

2.2 Consent Management Solutions

Ambiguity about how to technically implement the principles of privacy law [7] led to heterogeneity in consent management solutions. In response, the Internet Advertising Bureau (IAB) – not to be confused with the Internet Architecture Board – developed the Transparency and Consent Framework (TCF), "the only GDPR consent solution built by the industry for the industry" [20]. The TCF standardizes and centralizes the storage of 'global' consent cookies. It is visualized in Figure 2. We describe this technical standard to illustrate what CMPs do, and also because it is implemented by many but not all of the CMPs we measure in later sections.

The first building block of the TCF is the definition of purposes and features that are shown to users. In TCF 1.0, purposes define reasons for collecting personal data, for example; personalization, ad selection, or usage analytics. Features on the other hand describe methods of data use that overlap multiple purposes, such as combination with offline sources. A full list of purposes and features can be found in Table A.1. Both must be disclosed to the users, but users are only given control over consenting to individual purposes.

The second building block of the TCF is the Global Vendor List (GVL), a master list of advertisers participating in the framework. The GVL is maintained by the IAB. Vendors declare the purposes for which they collect data and the features upon which they rely. They can also declare legitimate interest for specific purposes, which allows them to process personal data under the GDPR even if the

user does not consent. For each advertiser, the GVL contains; a name, a link to the advertiser's privacy policy, the feature and purpose ids consent is requested for, and the declared legitimate interests. Registered advertisers pay a yearly management fee of 1.200 €. Cookie prompts implementing the TCF often request consent for all advertisers in this list, even though the website does not have a business relationship with every vendor. If the list is updated with new vendors (or additional purposes), users are prompted with a new dialogue in order to obtain additional consent.

The third building block involves the *Consent Management Providers* implementing the TCF on publishers' websites. They provide the cookie prompt, store the user's choice as a browser cookie, and provide an API for advertisers to access this information. The IAB also maintains a public list for CMPs, which lists 150 participating providers as of May 2020 [19]. A website wishing to implement the TCF independently must become a CMP, otherwise they can out-source this to an existing CMP. In reality, a handful of CMPs dominate the market.

Beyond the technical standard, IAB Europe also governs the surrounding ecosystem. The legal terms used in consent dialogues, such as the purposes of data collection, are standardized in the TCF. Firms adopting the standard are expected to follow the defined policy and IAB Europe publicizes a tool to audit CMPs (but not vendors). We go on to provide evidence that the TCF is inconsistently implemented in practice and not at all in some cases, such as for CMPs targeting the US market.

3 MEASUREMENT APPROACH

We identify our items of interest (what we want to measure) in Section 3.1. We map the items to a set of indicators and measurement methods that collectively describe our methodology in Section 3.2. Finally, we assess the threats to reliability and validity of our methodology in Section 3.5.

3.1 Items of Interest

The following items (**I1–I7**) span the consent ecosystem, which is visualized in Figure 2. In particular, the red arrows and pipes with pressure gauges are the links in the ecosystem that we measure. We know little about the prevalence of CMPs on the web. This complicates generalizing results about CMPs from snapshot samples of the most popular websites with size in the order of thousands, as was done in previous work [32, 39]. In order to build a fuller picture of the consent ecosystem, we ask: how does CMP adoption vary according to website popularity (**I1**), and related, how has this changed over time and been influenced by developments in privacy law (**I2**).

The third item of interest relates to publisher behaviour: to what extent do websites customize the embedded CMP (**I3**). Privacy laws describe how consent can be legally collected, violations of which have been studied in [32, 39]. The responsibility for such violations is far from clear when a website embeds a CMP, which is especially true when the CMP allows the website to customize the embedded consent dialogue.

Turning to vendors processing personal data, there are many reasons why a vendor might do so. Obtaining consent is not the only lawful basis for data processing. The fourth and fifth items are; why are vendors collecting personal data (**I4**), and what is their legal basis for doing so (**I5**).

One aspect that has not been considered in existing research is the additional effort required to reject data processing compared to accepting it. In most experiments, artificial dialogues are pre-installed on the subject's machine or loaded from a single source. In practice, users may already be habituated to the standardized CMP dialogs, but dialogs may need to send consent decisions to multiple vendors which incurs additional waiting time. This motivates our items at the user-interface; how long does it take CMPs to distribute consent decisions (**I6**), and to what extent does the user's dialog interaction time vary depending on which privacy preferences are expressed (**I7**).

3.2 Measurement Methodology

Large-Scale Web Measurement. To measure the prevalence of consent prompts longitudinally, we analyze automated browser crawls recorded by the Netograph web measurement platform[1] described in Figure 3. Netograph was not built exclusively for this research project and exhibits some unique properties compared to existing methods. Most prominently, instead of sampling from a particular toplist at one point in time, our crawlers are constantly seeded with new URLs shared on social media platforms.

This approach is not a design choice made specifically for our research, but useful in our context as measurements are not limited to a domain's landing page (https://example.com/) but also cover arbitrary subsites (https://example.com/foo?bar). Recent work has shown that subsites show a significant different behavior and an increase of privacy-invasive techniques [55].

Netograph ingests a live feed of social media posts, extracts all URLs, and submits them into a capture queue. URLs are visited once within a couple of minutes after submission. Crawls are performed on virtual machines in US and EU data centers of a large public cloud provider. 50% of crawls are done from within the EU, each URL is assigned randomly. Websites are opened using Google Chrome on Linux with its current default user agent[2], a desktop resolution of 1024×800, and `en-US` as the preferred browser language. All other settings are set to their defaults: Third party cookies are allowed, the "Do not Track" HTTP header is not set, and Flash is disabled. Due to the large volume of URLs, Netograph crawls with relatively aggressive timeouts, which are discussed further in Section 3.5.

For every capture, Netograph collects the following data points using custom browser instrumentation. First, HTTP headers are logged for all requests and responses. Additionally, connection-related metadata such as IP addresses and TLS certificate chains are stored. For every domain in a capture, its relation to the main page, all cookies, IndexedDB, LocalStorage, SessionStorage and WebSQL records are saved. Finally, a screenshot of the visible area (without scrolling) is taken. Netograph does not store page contents due to storage constraints. All crawl data is stored in a central database, which can be queried using a custom API. As of May 2020, this database stores 161,214,215 captures or about 23 billion HTTP requests.

[1] https://netograph.io/

[2] Currently `Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.61 Safari/537.36`.

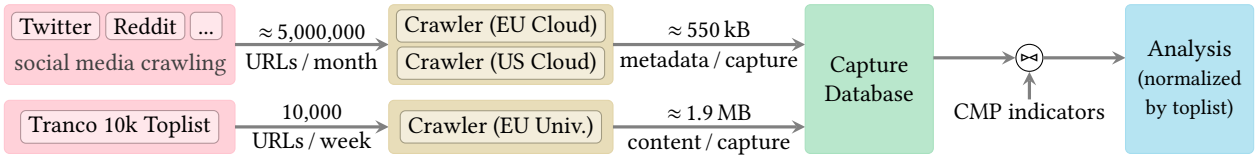

Figure 3: The Netograph measurement platform collects a realtime stream of URLs shared on social media and crawls them using Google Chrome. Custom browser instrumentation extracts metadata such as HTTP requests and cookies. We match captures with CMP indicators and use the Tranco toplist to normalize website popularity.

Toplist-Based Web Measurement. To make comparisons with related work, we have set up an additional Netograph-based crawling infrastructure for this study based on an internet toplist. In our analysis, we use the top 10k entries from the Tranco list created on 30 January 2020[3], which aggregates the ranks from the lists provided by Alexa, Cisco Umbrella, Majestic, and Quantcast [44]. This sample size is in the order of magnitude of previous studies (see # domains in Figure 1).

We first converted the Tranco list of domains to a list of URLs that can be crawled. For each *domain*, we attempted to establish a TLS connection to www.*domain* on port 443 and validate the certificate hostname using Mozilla's trust store. If the certificate is valid, we used https://www.*domain*/ as the seed URL for crawls. Otherwise, we attempted to open a TCP connection on port 80 and used http://www.*domain*/ on success. If this also failed, we used http://*domain*/ as the seed URL. We repeated this process three times over a week in order to catch temporarily unavailable domains.

Next, we crawled every URL in the toplist six times in immediate succession: First, we visited the website from a European university network using our crawler's default configuration. Second, we repeated this capture with an extended timeout. Third and fourth, we also captured with both German and British English as the preferred browser language. Finally, we submitted the same URLs to Netograph's task queues in the US and EU cloud as a control group. We retried all unsuccessful captures three times over the span of a week.

For all toplist crawls, we additionally stored the browser's DOM tree including the computed CSS styles. We also recorded a full-page screenshot (including scrolling). These extended features are not stored for the social media dataset due to their storage requirements.

*Prevalence and Customization of CMPs (**I1–I3**).* In the second part of our analysis, we measure the prevalence of CMPs using our crawl data. This involves extracting the final effective second-level domain (by which we want to count), detecting the CMP in use, and interpolating missing data. For this analysis we restrict ourselves to six CMPs: The five major players already identified by Nouwens et al. [39] and LiveRamp, a new entrant that launched in December 2019.

We measure the market share of CMPs by determining the number of domains they are active on. As about 11% of all crawls include top-level domain redirects, taking the domain from a seed URL would be imprecise. Instead, we extract the domain from the final website address as it would be shown in the browser's address bar. We normalize this domain to the effective second-level domain using the Public Suffix List [13], which contains all suffixes under which internet users can directly register names. For example, a capture may start with https://tinyurl.com/... as a seed URL, which redirects to https://foo.example.github.io/..., which we normalize to example.github.io.

To determine the CMP in use, we inspected the behavior of the six CMPs under study and created fingerprints for each CMP based on their HTTP request patterns, CSS selectors, and extracted text. For each CMP, we first recorded the network traffic of multiple websites where it was embedded and consulted the documentation provided by the CMP. Second, we assembled multiple fingerprints of varying specificity (for example, from concrete URLs to second-level domains) using manual analysis. To make sure that we did not miss any CMP dialogs, we searched for the GDPR phrases listed in [11] in our toplist crawls. We then checked the screenshots from our toplist crawls and discarded all fingerprints that yield false-positives. Finally, we verified that the remaining fingerprints work accurately for historic data using Netograph's captured screenshots. Using this approach, we were able to identify a unique hostname for each consent dialog framework as a robust indicator. For example, even though OneTrust deploys very different dialog designs with no shared JavaScript code or CSS classes, all of them perform HTTP requests to cdn.cookielaw.org on page load. We list our synthesized indicators in Table A.2 for reproducibility.

Finally, we also need to take into account that the sampling frequency of a domain is not fixed in our main dataset as the crawler is seeded from social media posts only. Consequently, we may not see less popular domains for prolonged periods. We account for this in two ways. First, we interpolate missing observation periods if both boundary measurements are classified equally. For example, if we observed Quantcast on example.com a month ago and observe it again today, we assume that example.com kept using Quantcast as their CMP throughout this period. If the boundary measurements disagree, we do not assume the presence of the CMP in the intermediate period. Second, we account for the fact that our measurements are right-censored by fading out the presence of a CMP after 30 days if no new measurements have been made yet. For example, if a website was last measured a week before our analysis, we assume that they still use the same CMP; if the last measurement was made on February 1st, we assume no CMP presence as of March 1st. Finally, as we crawl with a fixed sampling frequency for our toplist-based measurements, we do not need to interpolate for this dataset.

[3] Available at https://tranco-list.eu/list/K8JW.

*Ad-Tech Vendor Behavior (**I4–I5**).* Recall that Ad-tech vendors need to declare in the TCF for which data processing purposes they either request consent or claim legitimate interest. To assess the behavior of vendors, we systematically analyzed previous versions of the GVL and inspected them for longitudinal changes. In particular, we measure every instance when an Ad-tech vendor joins or leaves the GVL, claims a new purpose falls under legitimate interest, begins requesting consent for a new purpose, stops claiming either, or changes from collecting consent to claiming legitimate interest or the other way round.

*Time to Consent (**I6–I7**).* An aspect that has not been studied in the literature is the relative time taken to express different consent preferences. We aim to quantify this by embedding the dialogues offered by two leading CMPs, namely Quantcast and TrustArc. Using real dialogues in a field experiment improves ecological validity relative to studies using dialogues developed by researchers that result in a very different *feel* for the participants who are not browsing *normally*.

First, we measured how a seemingly small user interface change impacts the time it takes users to make a positive or negative consent decision. We embedded Quantcast's CMP dialog on a popular website on the public internet for a short period of time in two configurations: One with an explicit "Reject" button and one that included a "More Options" at the same position which would then lead to a reject button (see Figures A.1 and A.2). This design is motivated by the French data protection authority's guidelines, which demand a real choice between accepting or refusing cookies presented at the same level [10]. All other dialog settings were left to the default values: The consent prompt was shown as a modal dialog in the center of the screen, consent for all vendors on the GVL was requested, the "Accept" button was colored more prominently, and the dialog was only shown to visitors from the EU. We then measured the page load time (`DOMContentLoaded`), the time the dialog appeared (`__cmp('ping',...)`[4]), and the time it was closed as well as the user's consent decision (`__cmp('getConsentData',...)`). We also checked for the existence of already existing global consent cookies by manually fetching https://api.quantcast.mgr.consensu.org/CookieAccess, which returns the users's existing Quantcast TCF cookie. Repeated visitors will not be counted as the CMP stores the first consent decision and no additional dialogs will be shown.

Second, we noticed that some CMP dialogs require extended processing time if users decide to opt out. For example, TrustArc consent prompts disappear immediately if one accepts cookies, but otherwise make the user wait for prolonged periods while opt-out requests are being sent to a hodgepodge of third parties. In our testing, opting out required users to wait tens of seconds, which could be skipped at any time by giving consent. To make sure that these observations were not a fluke, we repeatedly visited a website embedding the TrustArc dialog, automated the opt-out process with a custom Google Chrome extension, and collected all HTTP requests and timings.

[4]The `__cmp()` function is standardized as part of the IAB's Transparency & Consent Framework, see Matte et al. [32].

3.3 Research Ethics

Our time-to-consent measurements were conducted on a website with real users, which raises ethical concerns as we did not ask for consent prior to measuring their interactions with consent notices. We did so to ensure non-biased results, which is supported by previous research on consent dialogs [56]. We ensured that we did not harm website visitors and their privacy. We address privacy issues by data minimization, i.e. we only collected a user's consent decision and the timings described in Section 3.2. The timings for a single page visit are linked using a random non-persistent id generated on page load. We do no create or store any persistent identifiers. While we believe that the second dialog design may not fulfill the requirements of the GDPR, the website we ran our experiments on did not perform any personal data collection irrespective of the user's consent decision.

3.4 Data Sources

Recall that Netograph's web crawlers are seeded with URLs posted on social media. More specifically, we ingest all URLs shared on Reddit and 1% of public Tweets using Twitter's sample feed[5]. Note that this does not mean we see 1% of URLs: each popular URL has multiple chances to be spotted in the sample feed as it is re-shared and retweeted. So in effect our URL sample skews heavily towards popular URLs. Overall, Twitter accounts for 80% of all URLs. We skip a URL if we have captured the same domain in the last hour or the precise URL in the last 48 hours. This applies to about 40% of all submitted URLs. Our records span March 2018–September 2020, starting before the inception of GDPR and also covering the introduction of CCPA.

To track the development of the global vendor list, we systematically downloaded all 215 previously published versions of the GVL from https://vendorlist.consensu.org/v*XXX*/vendor-list.json and verified their accuracy using the Internet Wayback Machine. Likewise, we collected the change history of Quantcast's consent dialog in the same way.

To measure how long it takes for users to make a consent decision, we embedded Quantcast's CMP dialog and our collection script on mitmproxy.org for a short period of time in May 2020. We logged about 120,000 timestamps. Importantly for generalizing, the website we hosted our experiment on caters to a very technical and privacy-concious audience.

For our second timing experiment, we measured the raw waiting time (not including user interaction) it takes to reject all tracking on forbes.com's TrustArc consent dialog. Measurements were performed hourly for two weeks in May 2020. These measurements were made from a European university as the vantage point.

The relationship between our items of interest, data sources, and vantage points is summarized again in Appendix A.4.

3.5 Reliability and Validity

Social Media Sample Bias. While existing research is mostly based on the Alexa and Tranco toplists, our measurement platform is seeded using URLs obtained from social media posts. An obvious issue with this setup is that URLs shared on social media are not a representative sample of the internet. One would reasonably expect

[5]https://developer.twitter.com/en/docs/labs/sampled-stream/overview

Location	US ☁	EU ☁	EU University			
User Agent	🇺🇸	🇺🇸	🇺🇸	🇺🇸	🇬🇧	🇩🇪
Timing				⏲	⏲	⏲
OneTrust	341	368	403	412	412	414
Quantcast	173	207	225	229	230	233
TrustArc	107	118	152	157	154	156
Cookiebot	92	97	96	98	99	99
LiveRamp	8	9	14	14	14	14
Crownpeak	8	8	8	9	9	9
∑	729	807	898	919	918	925
Coverage	79%	87%	97%	99%	99%	100%

Table 1: Occurence of CMPs on websites in the Tranco 10k measured from different vantage points.

YouTube videos to be shared more than mastercard.com. Hence our sample exhibits a different coverage error than typical toplist-based studies, which are not representative of the internet either. Additionally, our choice of social media data feeds is heavily skewed towards Western culture. We rectify this bias in part by grouping captures by their effective second-level domain. In other words, popular domains have a higher sampling frequency in our dataset, but equal weight.

Missing Data. Another threat to validity is that some domains in the toplist have never been shared on social media. This affects 1076 domains in the Tranco 10k list. Of these 1076 domains, 315 were not reachable via HTTP or HTTPS at all in our toplist measurements, 4 did not return a valid HTTP response and 70 returned an HTTP error status code. 192 domains redirected to another domain and were counted as the redirect target. The overwhelming majority (> 90%) of the remaining 495 domains can be considered internet infrastructure that is not directly accessed by users, such as CDNs.

Subsites. In contrast to previous research, we crawl not only a domain's landing page but also arbitrary subsites given by the seed URLs. This increases the reliability of our results as it allows us to detect CMPs that are only present on specific subdomains or subsites. However, we also encounter individual pages that do not include a CMP. For example, some websites do not embed any external scripts on their privacy policy page. As a simple heuristic, we classify a website as using a CMP if the CMP is included in at least every third capture. For 99.8% of all domains, the daily share of CMP captures is either consistently below 5% or above 95%.

The remaining 0.2% of websites include a small set of larger websites which change their behavior depending on the user's location, for example by complying with CCPA in the US but responding with `HTTP 451 Unavailable For Legal Reasons` to European visitors.

Crawler Location. Netograph crawls all URLs from virtual machines rented from a large public cloud provider. Half of all captures are done from the EU and the US respectively. This matches the recommendations made by Van Eijk et al. [58] to perform crawls from both inside and outside the EU for cookie consent notices. As shown in Table 1, we observe significantly more CMP adoption when crawling from the EU. This observation matches Van Eijk et al.'s finding on vantage point difference and can be explained by websites that only embed a CMP for EU visitors. Still, many websites choose to always embed their CMP framework but configure it to only show consent dialogs to EU visitors.

However, we found that not only the originating country, but also the type of address space has a significant influence on measurement results. As shown in Table 1, the use of public cloud infrastructure makes us miss about 10% of all CMP dialogs in the Tranco 10k. We manually inspected the sites in question and found that this is predominantly caused by anti-bot interstitial pages offered by popular CDNs. In contrast to the vantage point, the choice of browser language settings did not have a significant effect on our web measurements.

Lastly, we re-iterate our overall point that longitudinal measurements matter for web privacy measurements: Looking at the same measurements in January 2020 (see Table A.3), we see that only 70% of CMP usage is visible in our measurements from the US. The rise in coverage can be explained by the increasing adoption of CCPA in recent months.

Crawler Timeouts. Due to the large volume of URLs, Netograph runs crawls with relatively aggressive timeouts. To determine if a page has finished loading, it looks at frame load events from Chrome, the timing of requests, an idle timeout of five seconds and a total page timeout of 45 seconds. We note that crawls are done with heavy CPU utilization and a comparison with captures from the desktop might not be apt. In any case, our approach differs from smaller toplist-based measurements, which can afford much more relaxed timeouts. We quantify this change in Table 1: The timeouts employed by our measurement platform make us miss about 2% of CMP usage.

Choice of Toplist. To determine website popularity, we used the Tranco toplist [44]. Tranco aggregates results from other lists such as the Alexa toplist, is hardened against manipulation, less susceptible to daily fluctuations, and emphasizes reproducibility by providing permanent citable references. This decision is on line with recent related work on cookie consent [32]. While Urban et al. adapt the suggestion in the Tranco paper to remove all websites with the same TLD+1 [55], we do not perform this in our case as services may vary in their behavior across TLDs. For example, `amazon.com` shows a different consent prompt than the EU version of `amazon.co.uk` as of May 2020. A much more important factor which previous work has not elaborated on is the choice of toplist size. We show in the next section that different toplist sizes yield significantly different results.

CMP Detection. We found our detection of CMPs to be robust despite heterogeneous CMP implementations on different websites. By looking at network traffic patterns we do not rely on any HTML or DOM parsing, which we found to be much more unreliable for analyses which we ultimately decided not to include in this paper. In particular, network patterns often allow us to detect the presence of CMPs even if the website's CMP configuration does not trigger a dialog, for example because we visit a EU-centric website from the US or vice-versa. However, we acknowledge that

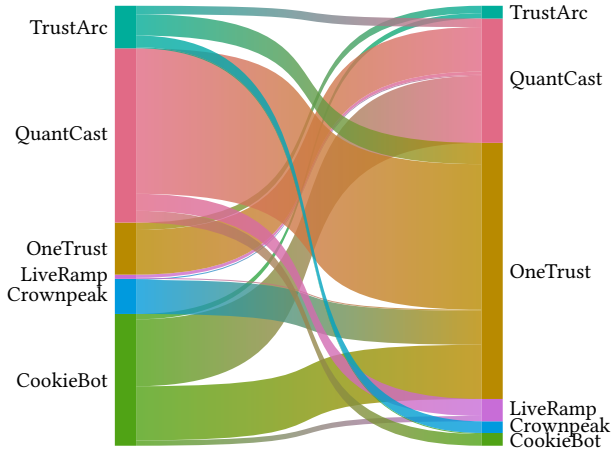

Figure 4: Sankey diagram of 414 CMP switches in the observation period (April 2018 to September 2020)

our detection accuracy and robustness is difficult to quantify. We have manually evaluated patterns on other candidate domains, patterns on specific HTTP requests, patterns on CSS selectors, and patterns on extracted text to make sure that we do not miss any CMP implementations. Additionally, we have used the Internet Wayback Machine to validate that our patterns match correctly on historic data. The only exception to this is a two-day period in July 2018 when Quantcast embedded parts of their CMP script for all customers of their analytics service, a different line of the firm's business. We manually exclude this outlier in our calculations. We overcount if a website includes more than one CMP, but this only affects 0.01% of all captures.

4 RESULTS

This section is structured according to which part of the ecosystem we are focusing on; websites and CMPs in Section 4.1, vendors in Section 4.2, and the user-interface in Section 4.3

4.1 Measuring CMP Adoption

Figure 5 shows how CMP adoption varies across the Tranco top million sites. The y-axis shows the percentage of firms embedding each CMP provider in the toplist with size corresponding to the x-axis. None of the largest websites embed the CMPs under consideration, likely because they have the in-house expertise to implement their own consent management solution. Speaking to (**I1**), CMP adoption is most prevalent among the $50 - 10,000^{\text{th}}$ websites, especially in the top $1,000 - 5,000^{\text{th}}$ sites. Adoption tails off slowly but never vanishes.

Interestingly, we see that different firms penetrate different sections of the market. For example, more of the top 100 sites embed Quantcast than the other CMP providers combined. However, OneTrust has the most customers among the $500 - 50,000^{\text{th}}$ sites, although Quantcast are more commonly adopted in the long tail.

Figure 6 shows how this has varied over time (**I2**). Laws like GDPR and CCPA coming into effect were significant drivers in CMP adoption, which suggests consent management solutions are more about regulatory compliance than improving user experience. However, events relevant to privacy law like fines or regulatory guidance do not affect adoption. Quantcast's solution is targeted at GDPR and they achieved market dominance early on, but their market growth slowed and was unaffected by the CCPA coming into effect. In contrast, OneTrust became the market leader by offering a flexible solution that could be tailored to the requirements of the CCPA. This can be seen in the share of sites with a EU+UK TLD for each CMP (Quantcast at 38.3% and OneTrust with 16.3%).

Our longitudinal approach can detect when websites change CMPs. Figure 4 describes the resulting dynamics. Quantcast and OneTrust both win and lose websites to each other. However, the true loser of inter-CMP competition is Cookiebot who have lost an order of magnitude more websites than they gained. The appendix contains further longitudinal insights by showing the CMP marketshare in January 2019, January 2020, and September 2020 (respectively Figure A.4, A.5 and A.6). These three figures show how OneTrust over-hauled the early market dominance established by Quantcast.

We now turn to how publishers customize consent solutions (**I3**). CMPs differ in how much customizability they extend to publishers, we classify this into *closed customization* in which the publisher may choose between finitely many options, and *open customization* in which the publisher can choose infinitely many, such as via free-text fields. In addition, *publisher customization* occurs when the website implements consent management related functionality beyond that offered by the CMP. We characterize the observed customization for the three largest CMPs to illustrate the ways in which this varies. All reported statistics are based on our measurements from an EU university vantage point (see Table 1) where we have the browser's DOM tree and full page screenshots available for inspection.

Our sample includes 414 websites embedding OneTrust displaying a range of consent dialogues. The majority (61%) offer a conventional cookie banner with a 1-click accept button and a second button or link leading to a page with more information and fine-grained controls. Only 2.4% of the sites display a cookie banner containing an opt-out button with text like "Do Not Sell", "Reject/Manage Cookies", or "Deny All", although 40% of such banners require further clicks to confirm the opt-out. A minority (5.5%) of websites include a 'script banner' (cookie banners in all but name) with one "Accept" button and one "Reject/Manage Scripts" button. Rather than showing any banner, 7.5% of the websites in our sample included a link to cookie or privacy information in the website footer. The link text was some variant of "Do Not Sell", "California Privacy Rights", or "Privacy Policy" in 11, 15 and 4 websites respectively. Two of the latter showed cookie banners only when accessed from a US IP.

Quantcast's dialogues are more standardized. Barriers contain two buttons, the first of which allows the user to provide consent to the publisher and partners in one click. Closed customizability is offered as a choice between the second-button rejecting all or it leading to a second page with more-fine grained options. Of the 233 websites embedding Quantcast in our sample, 55% offer a 1-click reject all. The text on each button is an interesting example of open-customization and we find that 87% use some variation of "I agree/consent/accept", including non-English language translations. The publishers who do not (13%) use free-form texts including

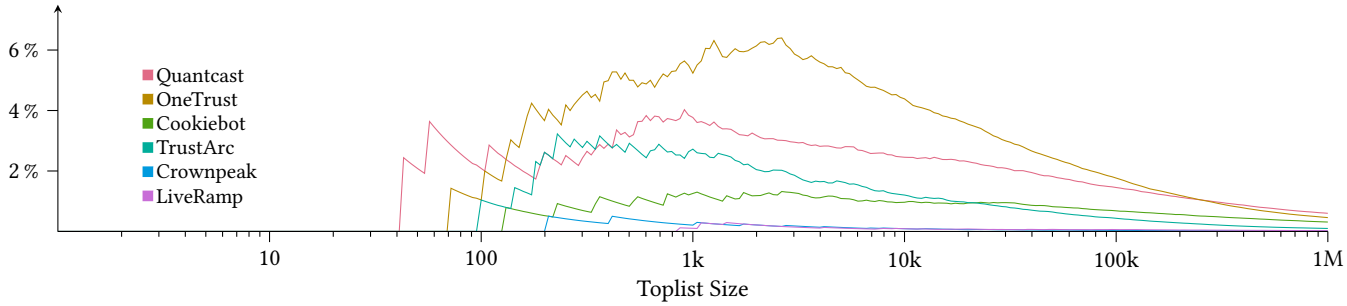

Figure 5: Cumulative CMP marketshare as a function of the toplist size (May 2020).

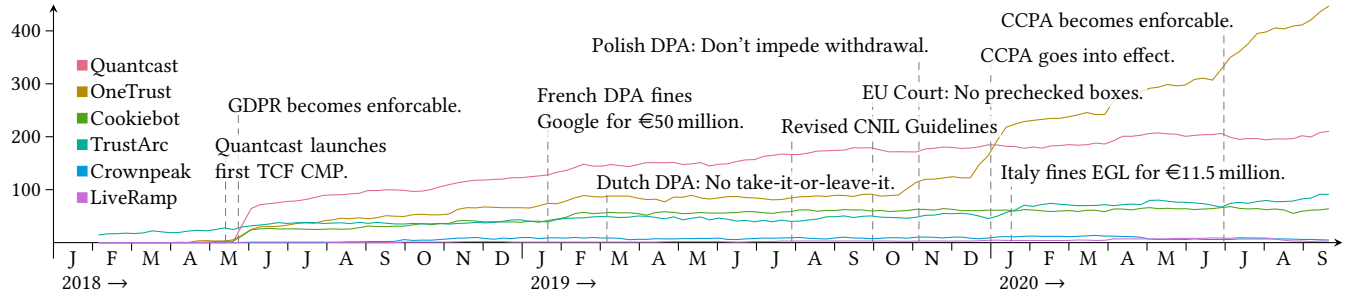

Figure 6: Number of websites in the Tranco 10k toplist that embed a CMP. We include a non-exhaustive timeline of events with relevance to the GDPR and the CCPA.

"Whatever", "Sounds good", and "Accept and move on" that may not qualify as affirmative consent.

TrustArc dialogues display more closed-customization in terms of button structure but have much less open-customization in terms of button wording. Of the 156 websites embedding TrustArc: 7% have a dialogue with a first-page button that instantly opts out; 12% have a first-page opt-out that must establish a connection with multiple partners (we measure the time to do so in Section 4.3); 44% include a first-page button that implies the user has autonomy; 31% have a link or button that does not imply the user has control; 4.4% hide their dialogue from EU IP addresses. TrustArc dialogues tend to define essential cookies for which there is no opt-out option. This, in combination with hiding dialogues from EU users, results from the product being tailored to the CCPA.

Finally, we estimate that about 8% of websites use CMPs for their APIs only and design custom consent dialogues themselves. This form of publisher customization presents a very practical problem: while these websites collect a standardized form of consent, each website does so in their own unique way, which may or may not comply with local legislation. As CMPs share consent across websites [60], this unreliable consent signal will then be re-used by other websites and third parties.

4.2 Measuring Third Party Vendors

The next two items of interest concern the purposes and lawful basis claimed by vendors for processing personal data. Using conventional methods, estimating how third-parties use personal data would require accessing and processing the privacy policy of each, which could be costly if repeated for longitudinal insights. In contrast, the IAB's standard allows us to measure this longitudinally for vendors on the Global Vendor List (GVL). In fact, the organization managing the GVL switched to weekly updates so we can detect all changes.

Figure 7 speaks to **I4**. It shows that both the size of the number of vendors and the reported purposes in the IAB's Global Vendor List have grown over time, with a sharp spike as GDPR came into effect. The first purpose, which allows vendors to collect and access personal data, is always the most popular. In Figure 7, it is difficult to track which movements are due to firms joining and which are due to an existing coalition member changing.

The changes made by existing members are summarized in Figure 8. This shows the surprising result that on net more vendors are now obtaining consent for purposes they used to claim as a legitimate interest than the other way round, which speaks to **I5**. This suggests that as time has passed, vendors on the GVL are obtaining more consent. The most activity regarding these changes took place around GDPR coming into effect, followed by another bout of activity in March and April 2020, possibly as vendors saw how GDPR was being enforced.

4.3 Measuring the User-Interface

Our results conclude with some findings regarding time costs related to consent dialogues. Our first item of interest here is the time it takes to send consent signals to multiple vendors (**I6**). We repeatedly measured the user's waiting time when they opt-out on a consent dialog provided by TrustArc and report the median numbers here. Figure 9 shows the opt-out process, which takes at

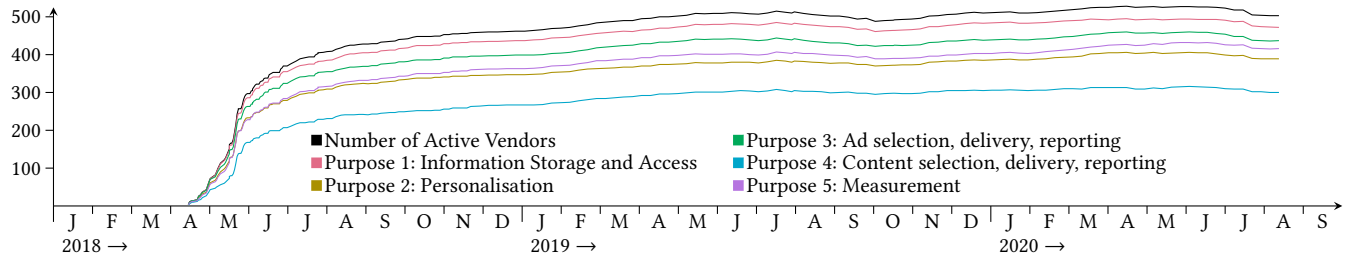

Figure 7: Reported purposes in the IAB's Global Vendor List

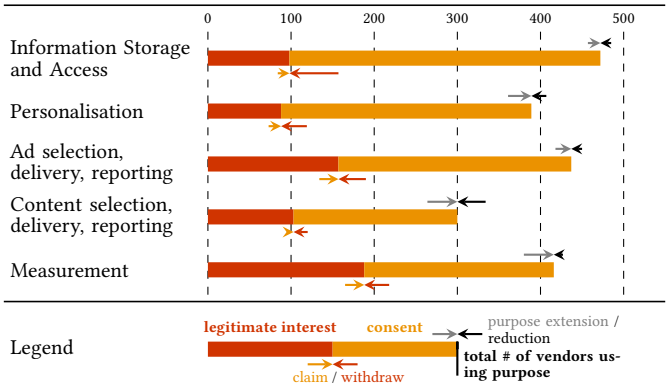

Figure 8: Purposes recorded in the IAB Global Vendor List

least 7 clicks and 34s to complete (not including user interaction). This delay results from sending opt-out requests to multiple third parties and additional JavaScript timeouts. Compared to accepting cookies, opting out causes an additional 279 HTTP(S) requests to 25 domains, which amounts to an additional 1.2 MB / 5.8 MB of data transfer (compressed / uncompressed). Thus in 12% of the websites embedding TrustArc (see Section 4.1), opting out is associated with a significant time and network cost for the user.

Second, we measured how the dialog interaction time varies depending on which privacy preferences are expressed (**I7**). Instead of using an artificial dialog design, we conducted a randomized experiment using Quantcast's real consent dialog in two different configurations further described in Section 3.2. In short, the first configuration included a direct reject button which was replaced with a "More Options" button in the second one (see Figures A.1–A.3). Section 4.1 showed that the first and second option were respectively used by 55% and 45% of websites embedding Quantcast dialogues. We exclude users who made no decision within the first three minutes after page load. In total, consent dialogs were shown to 2910 visitors from the EU (as per Quantcast's default configuration).

Our results are summarized in Figure 10: If Quantcast's dialog with a direct reject button is shown, it took the median user 3.2s to accept and 3.6s to deny consent. This difference is small but already statistically significant using a nonparametric test that is robust to skewed distibutions (Mann–Whitney $U(N_{\text{accept}} = 1344, N_{\text{reject}} = 279) = 166582$, $z = -2.93$, $p < 0.01$). If no direct reject button is shown, the median time it takes users to deny consent doubles to 6.7 seconds, which is highly significant ($U(N_{\text{accept}} = 1152, N_{\text{reject}} = 135) = 30494$, $z = -11.57$, $p < 0.001$). Additionally, the consent rate increases from 83% to 90%. In summary, we find that depending on the dialog design, the interaction time increases greatly for users who intend to opt out.

5 DISCUSSION

Section 5.1 discusses measurement issues like sampling and generalizing. Section 5.2 discusses the prevalence, significance, and future of consent management provision.

5.1 Methodological Implications

Social Media Sampling. Sampling URLs from social media posts is a novel approach through which we captured 161 million web pages from 4.2 million unique domains over a period of 2.5 years. This significantly exceeds the sample size and windows used in related work (see Figure 1). Building on recent approaches [55], subsite sampling is more tolerant to the many idiosyncrasies regarding how CMPs are embedded in the wild. At the same time, this sample is influenced by the social media websites' content filtering policies and–more importantly–heavily skewed towards the 'attention economy'. Such websites tend to be funded by collecting personal data, for which consent needs to be obtained. This bias is useful as we are more likely to sample websites that include CMPs.

We complement our social media crawling with a more traditional approach using the Tranco toplist. This means the proportions we estimate in Figures 5 and 6 are not affected by the social media sampling bias. However, top-lists are not representative of a meaningful population either, such as total web-page views or distinct sites visited by users. Given that both bottom-up sampling from social media posts and top-down sampling from toplists oversamples a certain population [49] with no ground-truth to adjust for it, using both approaches seems a defensible way forward.

Web Privacy Measurements. The notion that a web-page has a single set of observer-independent privacy features is dead [58]. We demonstrated that CMP adoption is influenced by local legislation and measurement results depend on vantage point (see Figure 1). Future empirical studies should take this into account and explain the implications for generalizing findings if only one vantage point is used.

Similarly, the occurrence of CMPs varies greatly depending on the toplist size (see Figure 5). From 4% in the Top 100, it reaches 13% in the Top 1k, and then falls in the long-tail down to 1.51% for

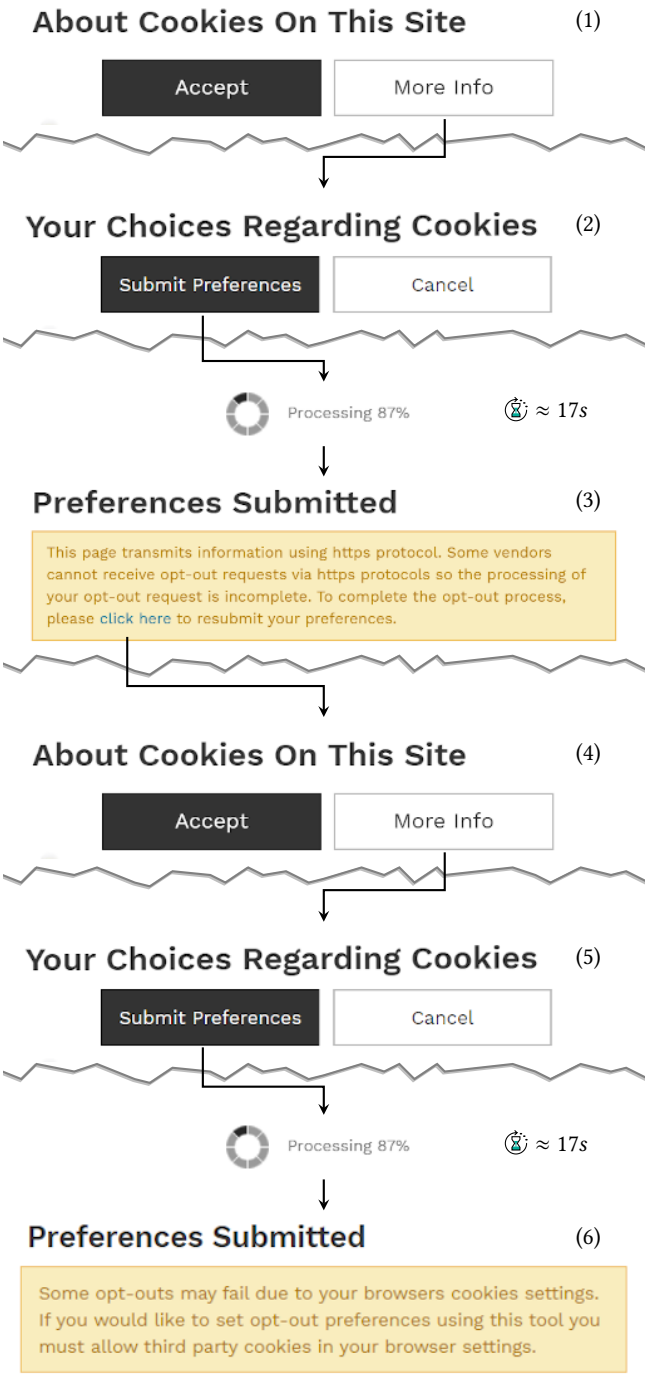

Figure 9: Training users to accept: Opting out on `forbes.com` takes at least 34 seconds (and seven clicks). Accepting cookies closes the dialog immediately.

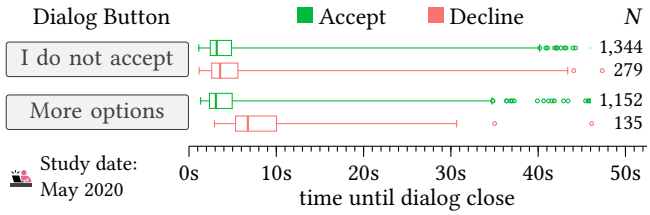

Figure 10: Randomized experiment with real CMP dialogs: depending on the dialog design, denying consent may take significantly longer than giving.

the Top 1M. These stark differences emphasize the importance of both sample size and choice of toplist from which it is drawn.

Web scraping can exploit common code structure across websites embedding CMPs. Such research designs can be scaled across the long tail of website popularity, which complements the qualitative analysis of tech giants [18]. However, it is not clear how such results generalize beyond websites employing CMPs. Similarly, we do not know how our results, based on six of the most popular CMPs, apply to niche CMPs[6] or websites self-implementing the TCF framework.

Measuring Ad-Tech Behavior. Given frameworks such as the TCF, the legal basis for third-party vendors can now be publicly queried and measured over time (see Figures 7 and 8) whereas previously this information was stored on corporate networks. However, we still cannot easily detect whether vendors adhere to self-declared policy.

5.2 Privacy Implications

Prevalence. We observed that CMPs are embedded in ever more websites over time and that privacy laws coming into effect caused spikes in adoption. The few times the GDPR was enforced had little observable effect (see Figure 6), although this could change if sanctions increase in frequency or significance. There is further churn between CMPs with Cookiebot functioning as a 'gateway CMP' that many websites adopt before migrating onto other CMPs (see Figure 4).

Significance. CMPs are standardizing privacy communications. The resulting legal terms, dialogue interface, and protocol for communicating with vendors should be seen as a de-facto standard, at least among that CMP's customers. Such standards were developed by self-interested private companies and not in the open bodies like the IETF or W3C, which raises questions about the politics of standards [26]. More positively, the consistent web interfaces provided by CMPs help researchers discover possible privacy violations at scale [32, 39], which mirrors researchers auditing compliance to credit card security standards [31, 46].

Beyond technical standards, CMPs can also influence social norms around privacy by herding websites. This can be seen in the linguistic shift from cookies to *scripts* that was only observed in 5.5% of the websites embedding OneTrust. This is likely a strategic move to escape the negative associations of cookies [54]. Herding

[6] Examples include Kochava, Adzerk CMP, and PreferenceManager.

may also strengthen the widely documented habituation effect in both privacy [5, 24, 59] and security notices [12].

Compliance with privacy laws drives CMP adoption, as evidenced by the spikes after the laws come into effect, and yet liability for violations is an open question. Quantcast maintain that "with great customizability comes great responsibility", which suggests they believe websites are liable for using terms like "whatever" as an affirmative signal of consent. Yet Quantcast offer dialogue functionality in which accepting takes 1-click while rejecting takes multiple, which is adopted by 45% of their customers, despite the French regulator's guidance against this practice [10].

Buttons allowing 1-click rejection are even rarer among websites embedding TrustArc (7%) and OneTrust (2.4%). The CMPs may know something its clients do not given trustarc.com implements an instant, 1-click reject all button. Disentangling whether these differences are driven by CMP business practices or pre-existing customer characteristics (e.g jurisdiction) can help prioritize regulatory interventions. The important role of intermediaries in (not) preventing abuse is an endearing lesson from information security economics [8, 35, 52], why would privacy economics be any different?

The specter of liability looms over vendors claiming a legitimate interest rather than obtaining consent [33]. For every purpose in the TCF, at least a fifth of the vendors claim they do not need to collect consent to process personal data (see Figure 8). More generally, one might ask why websites agree to collect consent for all of the Global Vendor List given there is no observed benefit to doing so [60].

The Future of Consent Management. If trends during the formation of the ecosystem continue, Figure 4 suggests that certain CMPs (Quantcast, OneTrust) will win market share from the others. A theoretical model predicts that sharing consent between the CMP's customers will create winner takes all dynamics leading to one global coalition [60]. In reality, jurisdictional boundaries will likely lead to multiple distinct coalitions given Quantcast and OneTrust appear to be establishing dominance in the EU+UK and the US respectively. However, users do not respect such jurisdictions. This will likely exacerbate the extent to which the web differs based on where the user appears to be located, which we observed at multiple points in this study.

The rise of CMPs should be seen as part of a wider process by which legal compliance shapes the internet. Liability for content shared on technology platforms provides another example [15] in light of a May 2020 executive order in the US. This represents a departure from utopian views of the Internet as a libertarian paradise [3]. One might begin to consider a *compliance layer* of the internet driven by the content and privacy policies of private firms as influenced by national laws. Before regulators demand measurements as evidence, the community should reflect on how to support auditing at scale, evidential standards, and surrounding ethical issues.

6 RELATED WORK

Returning to the piping metaphor of Figure 2, consent flows from a user's privacy preferences through a consent dialogue to the recipient of the consent signal and then on to third-parties. This section identifies related work at each interface, though none of the studies make measurements at as many interfaces as we do.

Qualitative research exploring privacy preferences of users informs internet design by, for example, identifying disparities between what users want and what happens online [4, 23, 40] or by highlighting the business value of obtaining explicit consent [61].

At the user-interface, lab experiments have consistently shown users can be shifted towards providing consent by changing framing [2, 5] and design choices, such as default settings [28, 30] and positioning [56]. Nouwens et al. [39] scrape post-GDPR UK websites to identify popular design choices and show that common practices like not having fine-grained controls on the first page increases propensity to consent. Our controlled experiment with real CMP dialogs on a public website complemented this body of work by showing users incur differing time costs based on the privacy preferences they express, highlighting how this punishes privacy aware users.

The next point of the consent flow concerns how consent dialogues interact with websites. Around 50% of the websites in [39] do not offer a 1-click opt out, which is confirmed by our samples of Quantcast websites. A dialogue or cookie banner may not even be shown. Degeling et al. [11] showed that 62 % of sampled European websites displayed cookie prompts right after GDPR came into effect in May 2018, up from 46 % in January 2018. However, these effects are not limited to Europe as websites in the US "approach cookie regulations similarly to the EU" [48], though this is not true of Chinese websites.

Turning to third-parties, research has predominantly focused on the extent of third-party tracking rather than how third-parties obtain consent (the final part of the consent flow). Iordanou et al. [22] introduce a methodology for measuring tracking at scale and show that the majority of tracking flows across European borders but, surprisingly, remains within the EU. Sørensen and Kosta [50] do not establish any change in the number of third-party trackers before and after GDPR, although they show that third-party tracking is more prevalent in private websites than public. Even after GDPR, Sanchez-Rola et al. [48] show that 90 % of sampled websites use cookies that could be used to identify users. Such results are hard to evaluate without more context. For example, a website needs to identify users who have not consented in order to not repeatedly present consent dialogues, which would violate the California Consumer Privacy Act.

Basing measurement on the TCF standard provides a way forward, Matte et al. [32] analyze sites using the TCF and find disparities between which preferences were communicated and which were stored as global cookies, which is more reliable evidence of a privacy violation. For example, 12 % of websites send the consent signal before the user even makes a choice and some even record the user's consent after an explicit opt-out. In a different study, the same authors build a legal argument that the purposes in the TCF are not specific or explicit enough "to be used as legally-compliant ones" [33] and measure which vendors claim these as a legitimate interest.

Finally, a theoretical work [60] considers the economic implications of CMPs forming 'consent coalitions' in which consent is shared across websites and vendors. Our measurements contradict their theoretical prediction about a 'global coalition', which does

not exist at present. The market will, however, further mature and our longitudinal results suggest a trend towards dominant CMPs in particular jurisdictions.

Considering our contribution to each aspect of online privacy in isolation obscures how our measurement approach allowed us to make longitudinal measurements across the entire consent ecosystem. Similar ecosystem wide measurements include those of: the advertising industry [14, 42, 43, 45]; online gaming [41]; VPN services [21, 25]; web communities [62, 63]; and web porn [57]. All of these studies, including ours, blend technical measurements with considerations around the economic and social factors influencing the agents in the ecosystem. Such studies provide a rigorous, empirical basis for how social scientists theorize about the impact of the Internet.

7 CONCLUSION

Recent years have seen the formation of a consent ecosystem through which websites and third-party vendors establish a legal basis for business models based on personal data. Our longitudinal approach tracks the rise of CMPs from less than 1% of the Tranco 10k toplist in February 2018 to almost 10% in September 2020 and we show that privacy laws (GDPR and CCPA) coming into effect caused spikes in adoption. We document inter-firm competition by which certain CMPs (e.g Cookiebot) bleed customers while others slowly establish dominance in a specific jurisdiction, such as Quantcast in the EU+UK or OneTrust in the US. This increasing market dominance allows private actors (often tied to the Ad-tech industry) to standardize the terms user consent to, the user-interface through which they do it, and also how it is shared with third-parties.

Although increasing market power is worrying, the same standardization opens up novel measurements opportunities. We tracked how third-party vendors justified their data processing activities, capturing changes over time like the shift towards obtaining consent. Similarly, we showed how the consent dialogues offered by CMPs impose a time cost on privacy aware users. These exact dialogues are used by the CMP's customers, which improved the ecological validity of our real-user study. More generally, regulators could exploit the structure provided by CMPs to audit privacy practices at scale.

ACKNOWLEDGEMENTS

We would like to thank Aldo Cortesi for his continuous support and the generous access to the Netograph API and capturing technology. We thank Tobias Kupek for his help with preparing figures. This work was co-funded by Archimedes Privatstiftung, Innsbruck. The second author is funded by the European Commission's call H2020-MSCA-IF-2019 under grant number 894700.

REFERENCES

[1] 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 (CCS '14)*. ACM, 674–689. https://doi.org/10.1145/2660267.2660347

[2] Idris Adjerid, Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2013. Sleights of Privacy: Framing, Disclosures, and the Limits of Transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. ACM, Article 9, 11 pages. https://doi.org/10.1145/2501604.2501613

[3] John Perry Barlow. 1996. A Declaration of the Independence of Cyberspace.

[4] Bettina Berendt, Oliver Günther, and Sarah Spiekermann. 2005. Privacy in E-Commerce: Stated Preferences vs. Actual Behavior. *Commun. ACM* 48, 4 (April 2005), 101–106. https://doi.org/10.1145/1053291.1053295

[5] Rainer Böhme and Stefan Köpsell. 2010. Trained to Accept? A Field Experiment on Consent Dialogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 2403–2406. https://doi.org/10.1145/1753326.1753689

[6] Joseph Bonneau and Sören Preibusch. 2009. The Privacy Jungle: On the Market for Data Protection in Social Networks. In *8th Annual Workshop on the Economics of Information Security, WEIS*. https://doi.org/10.1007/978-1-4419-6967-5_8

[7] Aaron Ceross and Andrew Simpson. 2018. Rethinking the Proposition of Privacy Engineering. In *Proceedings of the New Security Paradigms Workshop (NSPW '18)*. ACM, 89–102. https://doi.org/10.1145/3285002.3285006

[8] Richard Clayton, Tyler Moore, and Nicolas Christin. 2015. Concentrating Correctly on Cybercrime Concentration. In *14th Annual Workshop on the Economics of Information Security, WEIS*.

[9] Lorrie Faith Cranor. 2003. P3P: Making Privacy Policies More Useful. *IEEE Security and Privacy* 1, 6 (2003), 50–55. https://doi.org/10.1109/MSECP.2003.1253568

[10] Commission Nationale de l'Informatique et des Libertés (CNIL). 2019. *Guidelines on cookies and tracking devices*. https://www.cnil.fr/en/cookies-and-other-tracking-devices-cnil-publishes-new-guidelines

[11] 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)*. The Internet Society. https://doi.org/10.14722/ndss.2019.23378

[12] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, 1065–1074. https://doi.org/10.1145/1357054.1357219

[13] Mozilla Foundation. 2007–2020. *Public Suffix List*. https://publicsuffix.org/

[14] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. 2013. Follow the Money: Understanding Economics of Online Aggregation and Advertising. In *Proceedings of the 2013 Conference on Internet Measurement Conference (IMC '13)*. ACM, 141–148. https://doi.org/10.1145/2504730.2504768

[15] Tarleton Gillespie. 2010. The politics of 'platforms'. *New Media & Society* 12, 3 (2010), 347–364. https://doi.org/10.1177/1461444809342738

[16] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman M. Sadeh, and Florian Schaub. 2019. An Empirical Analysis of Data Deletion and Opt-Out Choices on 150 Websites. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS '19)*. USENIX. https://www.usenix.org/conference/soups2019/presentation/habib

[17] Elizabeth Liz Harding, Jarno J Vanto, Reece Clark, L Hannah Ji, and Sara C Ainsworth. 2019. Understanding the scope and impact of the California Consumer Privacy Act of 2018. *Journal of Data Protection & Privacy* 2, 3 (2019), 234–253.

[18] Soheil Human and Florian Cech. 2021. A Human-centric Perspective on Digital Consenting: The Case of GAFAM. In *Human Centred Intelligent Systems*. Springer, 139–159. https://doi.org/10.1007/978-981-15-5784-2_12

[19] IAB Europe. 2020. *CMP List*. https://iabeurope.eu/cmp-list/

[20] IAB Europe. 2020. *What is the Transparency and Consent Framework (TCF)?* https://iabeurope.eu/transparency-consent-framework/

[21] Muhammad Ikram, Narseo Vallina-Rodriguez, Suranga Seneviratne, Mohamed Ali Kaafar, and Vern Paxson. 2016. An Analysis of the Privacy and Security Risks of Android VPN Permission-Enabled Apps. In *Proceedings of the 2016 Internet Measurement Conference (IMC '16)*. ACM, 349–364. https://doi.org/10.1145/2987443.2987471

[22] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. 2018. Tracing Cross Border Web Tracking. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, 329–342. https://doi.org/10.1145/3278532.3278561

[23] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara B. Kiesler. 2015. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS '15)*. USENIX, 39–52. https://www.usenix.org/conference/soups2015/proceedings/presentation/kang

[24] Farzaneh Karegar, John Sören Pettersson, and Simone Fischer-Hübner. 2020. The Dilemma of User Engagement in Privacy Notices: Effects of Interaction Modes and Habituation on User Attention. *ACM Transactions on Privacy and Security* 23, 1 (2020), 5:1–5:38. https://doi.org/10.1145/3372296

[25] Mohammad Taha Khan, Joe DeBlasio, Geoffrey M. Voelker, Alex C. Snoeren, Chris Kanich, and Narseo Vallina-Rodriguez. 2018. An Empirical Analysis of the Commercial VPN Ecosystem. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, 443–456. https://doi.org/10.1145/3278532.3278570

[26] David M. Kristol. 2001. HTTP Cookies: Standards, privacy, and politics. *ACM Transactions on Internet Technology* 1, 2 (2001), 151–198. https://doi.org/10.1145/502152.502153

[27] Ponnurangam Kumaraguru, Lorrie Cranor, Jorge Lobo, and Seraphin Calo. 2007. A Survey of Privacy Policy Languages. In *Workshop on Usable IT Security Management: 3rd Symposium on Usable Privacy and Security, ACM (USM '07)*.

[28] Yee-Lin Lai and Kai-Lung Hui. 2006. Internet Opt-in and Opt-out: Investigating the Roles of Frames, Defaults and Privacy Concerns. In *Proceedings of the 2006 ACM SIGMIS CPR Conference on Computer Personnel Research (SIGMIS CPR '06)*. ACM, 253–263. https://doi.org/10.1145/1125170.1125230

[29] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. 2020. Browser Fingerprinting: A Survey. *ACM Transactions on the Web* 14, 2, Article 8 (April 2020), 33 pages. https://doi.org/10.1145/3386040

[30] Dominique Machuletz and Rainer Böhme. 2020. Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs after GDPR. *Proceedings on Privacy Enhancing Technologies* 2, 481–498. https://doi.org/10.2478/popets-2020-0037

[31] Samin Yaseer Mahmud, Akhil Acharya, Benjamin Andow, William Enck, and Bradley Reaves. 2020. Cardpliance: PCI DSS Compliance of Android Applications. In *29th USENIX Security Symposium (USENIX '20)*. 1517–1533. https://www.usenix.org/conference/usenixsecurity20/presentation/mahmud

[32] 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*. IEEE, 791–809. https://doi.org/10.1109/SP40000.2020.00076

[33] Célestin Matte, Cristiana Santos, and Nataliia Bielova. 2020. Purposes in IAB Europe's TCF: which legal basis and how are they used by advertisers?. In *Annual Privacy Forum (APF 2020)*.

[34] Jonathan R. Mayer and John C. Mitchell. 2012. Third-Party Web Tracking: Policy and Technology. In *2012 IEEE Symposium on Security and Privacy*. IEEE, 413–427. https://doi.org/10.1109/SP.2012.47

[35] Damon McCoy, Hitesh Dharmdasani, Christian Kreibich, Geoffrey M. Voelker, and Stefan Savage. 2012. Priceless: The Role of Payments in Abuse-Advertised Goods. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS '12)*. ACM, 845–856. https://doi.org/10.1145/2382196.2382285

[36] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Journal of Law and Policy for the Information Society* 4 (2008), 543.

[37] Lynette I. Millett, Batya Friedman, and Edward Felten. 2001. Cookies and Web Browser Design: Toward Realizing Informed Consent Online. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. ACM, 46–52. https://doi.org/10.1145/365024.365034

[38] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2013. Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting. In *2013 IEEE Symposium on Security and Privacy*. IEEE, 541–555. https://doi.org/10.1109/SP.2013.43

[39] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-Ups and Demonstrating Their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–13. https://doi.org/10.1145/3313831.3376321

[40] Judith S. Olson, Jonathan Grudin, and Eric Horvitz. 2005. A Study of Preferences for Sharing and Privacy. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1985–1988. https://doi.org/10.1145/1056808.1057073

[41] Mark O'Neill, Elham Vaziripour, Justin Wu, and Daniel Zappala. 2016. Condensing Steam: Distilling the Diversity of Gamer Behavior. In *Proceedings of the 2016 Internet Measurement Conference (IMC '16)*. ACM, 81–95. https://doi.org/10.1145/2987443.2987489

[42] Michalis Pachilakis, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2019. No More Chasing Waterfalls: A Measurement Study of the Header Bidding Ad-Ecosystem. In *Proceedings of the Internet Measurement Conference (IMC '19)*. ACM, 280–293. https://doi.org/10.1145/3355369.3355582

[43] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. 2017. If You Are Not Paying for It, You Are the Product: How Much Do Advertisers Pay to Reach You?. In *Proceedings of the 2017 Internet Measurement Conference (IMC '17)*. ACM, 142–156. https://doi.org/10.1145/3131365.3131397

[44] Victor Le Pochat, Tom van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *26th Annual Network and Distributed System Security Symposium (NDSS '19)*. The Internet Society. https://doi.org/10.14722/ndss.2019.23386

[45] 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 (IMC '15)*. ACM, 93–106. https://doi.org/10.1145/2815675.2815705

[46] Sazzadur Rahaman, Gang Wang, and Danfeng (Daphne) Yao. 2019. Security Certification in Payment Card Industry: Testbeds, Measurements, and Recommendations. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. ACM, 481–498. https://doi.org/10.1145/3319535.3363195

[47] Hana Ross. 2017. Data subject consent: How will the General Data Protection Regulation affect this? *Journal of Data Protection & Privacy* 1, 2 (2017), 146–155.

[48] 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 (Asia CCS '19)*. ACM, 340–351. https://doi.org/10.1145/3321705.3329806

[49] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, 478–493. https://doi.org/10.1145/3278532.3278574

[50] 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 (WWW '19)*. ACM, 1590–1600. https://doi.org/10.1145/3308558.3313524

[51] Sarah Spiekermann, Alessandro Acquisti, Rainer Böhme, and Kai Lung Hui. 2015. The challenges of personal data markets and privacy. *Electronic Markets* 25, 2 (2015), 161–167. https://doi.org/10.1007/s12525-015-0191-0

[52] Samaneh Tajalizadehkhoob, Tom Van Goethem, Maciej Korczyński, Arman Noroozian, Rainer Böhme, Tyler Moore, Wouter Joosen, and Michel van Eeten. 2017. Herding Vulnerable Cats: A Statistical Approach to Disentangle Joint Responsibility for Web Security in Shared Hosting. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*. ACM, 553–567. https://doi.org/10.1145/3133956.3133971

[53] 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. https://doi.org/10.2478/popets-2019-0023

[54] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*. ACM, Article 4, 15 pages. https://doi.org/10.1145/2335356.2335362

[55] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Beyond the Front Page: Measuring Third Party Dynamics in the Field. In *Proceedings of The Web Conference 2020 (WWW '20)*. ACM, 1275–1286. https://doi.org/10.1145/3366423.3380203

[56] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. ACM, 973–990. https://doi.org/10.1145/3319535.3354212

[57] 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)*. ACM, 245–258. https://doi.org/10.1145/3355369.3355583

[58] Rob Van Eijk, Hadi Asghari, Philipp Winter, and Arvind Narayanan. 2019. The Impact of User Location on Cookie Notices (Inside and Outside of the European Union). In *Workshop on Technology and Consumer Protection (ConPro'19)*.

[59] Anthony Vance, David Eargle, Jeffrey L. Jenkins, C. Brock Kirwan, and Bonnie Brinton Anderson. 2019. The Fog of Warnings: How Non-essential Notifications Blur with Security Warnings. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS '19)*. USENIX. https://www.usenix.org/conference/soups2019/presentation/vance

[60] Daniel W Woods and Rainer Böhme. 2020. The Commodification of Consent. In *20th Annual Workshop on the Economics of Information Security, WEIS*.

[61] Scott A Wright and Guang-Xin Xie. 2019. Perceived Privacy Violation: Exploring the Malleability of Privacy Expectations. *Journal of Business Ethics* 156, 1 (2019), 123–140. https://doi.org/10.1007/s10551-017-3553-z

[62] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18)*. ACM, 188–202. https://doi.org/10.1145/3278532.3278550

[63] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other through the Lens of Mainstream and Alternative News Sources. In *Proceedings of the 2017 Internet Measurement Conference (IMC '17)*. ACM, 405–417. https://doi.org/10.1145/3131365.3131390

A APPENDIX

Purposes Definitions	
1	**Information storage and access:** The storage of information, or access to information that is already stored, on your device such as advertising identifiers, device identifiers, cookies, and similar technologies.
2	**Personalisation.** The collection and processing of information about your use of this service to subsequently personalise advertising and/or content for you in other contexts, such as on other websites or apps, over time.
3	**Ad selection, delivery, reporting.** The collection of information, and combination with previously collected information, to select and deliver advertisements for you, and to measure the delivery and effectiveness of such advertisements.
4	**Content selection, delivery, reporting.** The collection of information, and combination with previously collected information, to select and deliver content for you, and to measure the delivery and effectiveness of such content.
5	**Measurement.** The collection of information about your use of the content, and combination with previously collected information, used to measure, understand, and report on your usage of the service.
Careful readers may note that "information storage and access" is not a purpose for personal data processing in itself, but an artifact of the obligations imposed by Article 5(3) of the ePrivacy Directive.	
Feature Definitions	
1	**Offline data matching.** Combining data from offline sources that were initially collected in other contexts with data collected online in support of one or more purposes.
2	**Device linking.** Processing data to link multiple devices that belong to the same user in support of one or more purposes.
3	**Precise geographic location data.** Collecting and supporting precise geographic location data in support of one or more purposes.

Table A.1: Purposes and features as defined in version 1 of the IAB's Trust and Consent Framework.

CMP	Unique Hostname
OneTrust	cdn.cookielaw.org
Quantcast	quantcast.mgr.consensu.org
TrustArc	consent.trustarc.com
Cookiebot	consent.cookiebot.com
LiveRamp	cmp.choice.faktor.io
Crownpeak	iabmap.evidon.com

Table A.2: Hostnames used as an indicator for the presence of a CMP (see Section 3.2).

Location	US ☁	EU ☁	EU University		
User Agent	🇺🇸	🇺🇸	🇺🇸	🇬🇧	🇩🇪
Timing			⏱	⏱	⏱
OneTrust	263	306	344	339	342
Quantcast	151	192	222	220	221
TrustArc	102	110	170	168	168
Cookiebot	82	90	92	92	92
LiveRamp	6	6	10	10	10
Crownpeak	9	10	34	35	34
$\sum$	613	714	872	864	867
Coverage	70%	82%	100%	99%	99%

Table A.3: Occurence of CMPs measured in January 2020. Comparing this to the May 2020 data in Table 1, we see that a growing share of websites adapt CMPs outside the EU, likely prompted by non-EU regulations such as CCPA.

Item of Interest	Vantage Point	Dataset
I1 CMP Adoption (by rank)	US/EU Cloud	Social media URLs from Tranco 1M
I2 CMP Adoption (over time)	US/EU Cloud	Social media URLs from Tranco 10k
I3 Publisher Customization	EU University	Tranco 10k front pages
I4 Collection Purposes	–	IAB Global Vendor List
I5 Legal Basis for Collection	–	IAB Global Vendor List
I6 Cost to Opt-Out	EU University	Measurements for `forbes.com`
I7 User Behavior	Visitors from EU countries	User study hosted on `mitmproxy.org`

Table A.4: Overview of the vantage points and datasets used for each measurement (see Section 3.2). For the first two items of interest, each URL is randomly distributed to either a US or a EU cloud instance for crawling.

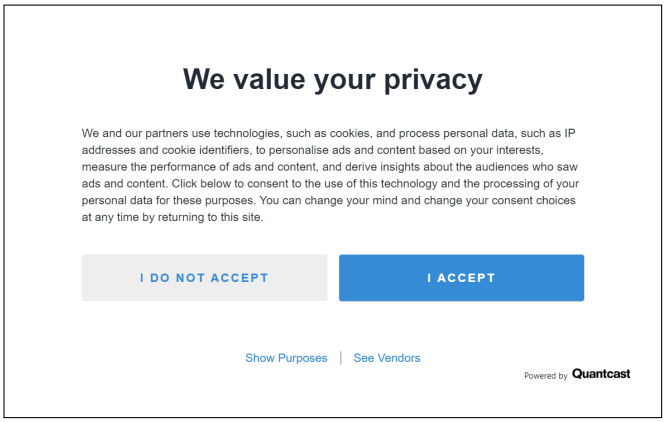

Figure A.1: Default version of Quantcast's consent dialog. The dialog is shown as a modal popup with a dark-gray background covering the rest of the page.

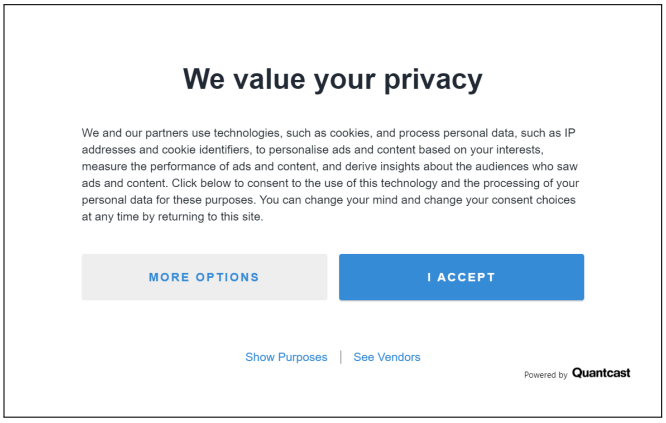

Figure A.2: Quantcast's consent dialog without direct reject option.

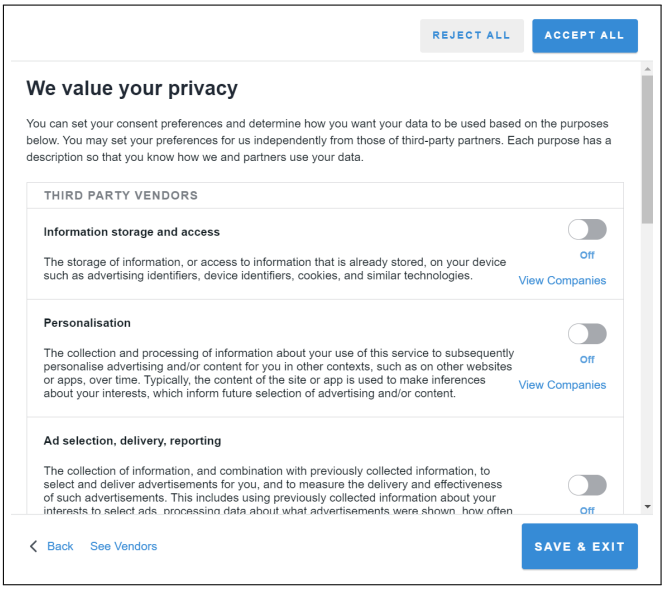

Figure A.3: Dialog shown to users after they click "More Options".

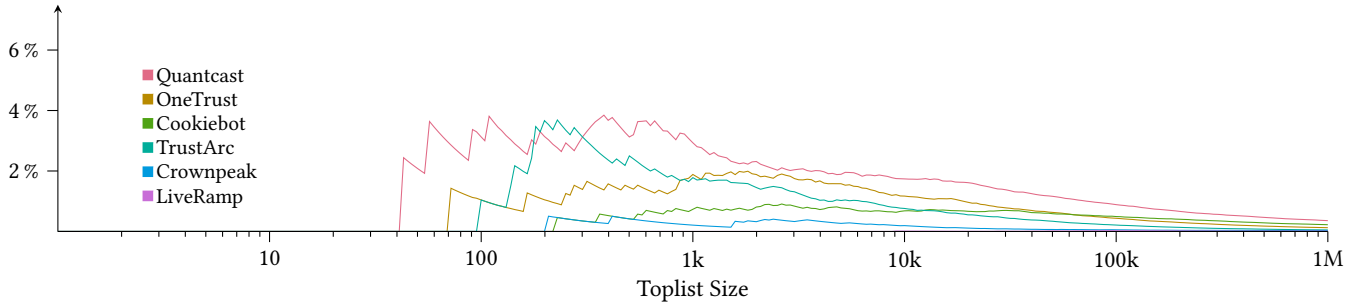

Figure A.4: CMP Marketshare (January 2019).

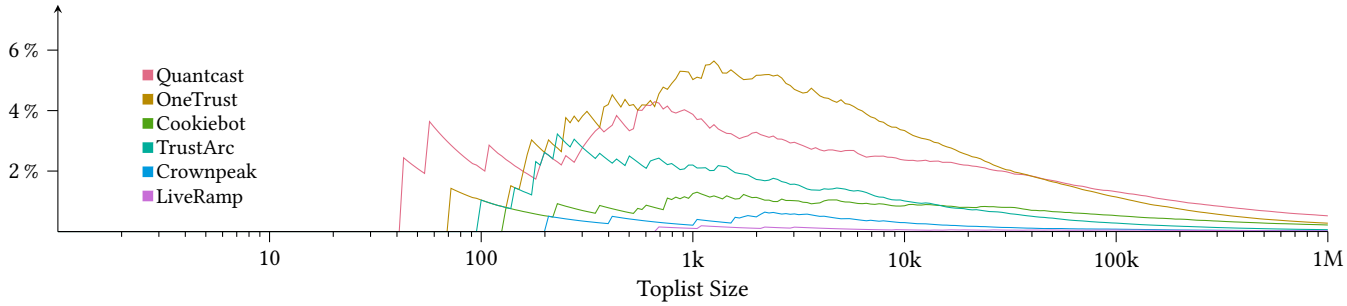

Figure A.5: CMP Marketshare (January 2020).

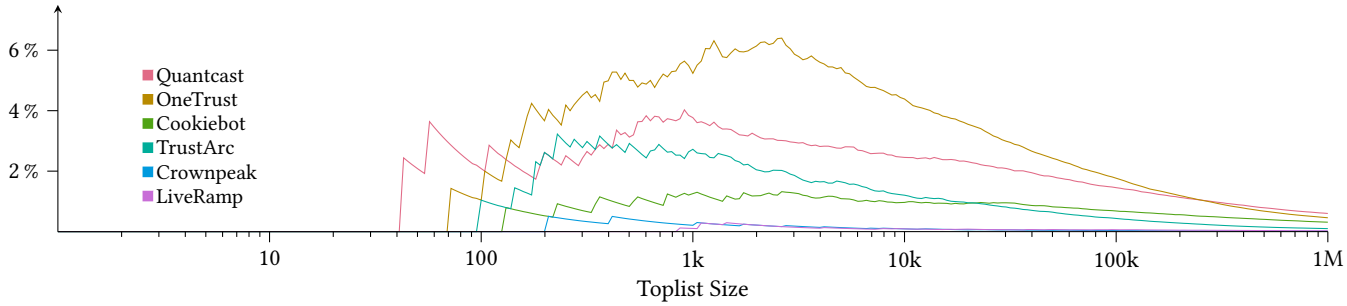

Figure A.6: Cumulative CMP marketshare as a function of the toplist size (May 2020). This is a repetition of Figure 5 included for better comparison.