sciendo

Proceedings on Privacy Enhancing Technologies ; 2022 (1):608–628

Maggie Van Nortwick and Christo Wilson

Setting the Bar Low: Are Websites Complying With the Minimum Requirements of the CCPA?

Abstract: On June 28, 2018, the California State Legislature passed the California Consumer Privacy Act (CCPA), arguably the most comprehensive piece of online privacy legislation in the United States. Online services covered by the CCPA are required to provide a hyperlink on their homepage with the text “Do Not Sell My Personal Information” (DNSMPI). The CCPA went into effect on January 1, 2020, a date that was chosen to give data collectors time to study the new law and bring themselves into compliance.

In this study, we begin the process of investigating whether websites are complying with the CCPA by focusing on DNSMPI links. Using longitudinal data crawled from the top 1M websites in the Tranco ranking, we examine which websites are including DNSMPI links, whether the websites without DNSMPI links are out of compliance with the law, whether websites are using geofences to dynamically hide DNSMPI links from non-Californians, how DNSMPI adoption has changed over time, and how websites are choosing to present DNSMPI links (e.g., in terms of font size, color, and placement). We argue that the answers to these questions are critical for spurring enforcement actions under the law, and helping to shape future privacy laws and regulations, e.g., rule making that will soon commence around the successor to the CCPA, known as the CPRA.

Keywords: CCPA, CPRA, Do Not Sell My Personal Information, Privacy Policies

DOI 10.2478/popets-2022-0030
Received 2021-05-31; revised 2021-09-15; accepted 2021-09-16.

1 Introduction

On June 28, 2018, the California State Legislature passed the California Consumer Privacy Act (CCPA) [27, 28], arguably the most comprehensive piece of online privacy legislation in the United States. The CCPA grants California residents new rights to: know what personal data is collected about them; know about and opt-out of the sale and sharing of personal data; access and request the deletion of collected data; and not be discriminated against for exercising their rights under the CCPA. Online services covered by the CCPA[1] are required to provide a hyperlink on their homepage with the text “Do Not Sell My Personal Information”[2] (which we abbreviate as DNSMPI) that leads to a webpage where people may engage their CCPA rights [22].

The CCPA went into effect on January 1, 2020, a date that was chosen to give data collectors (e.g., website and app developers, data brokers, online advertisers, etc.) time to study the new law and bring themselves into compliance. Additionally, the California Department of Justice (CA DoJ) engaged in an extensive rule-making process around the CCPA to clarify and strengthen its provisions, with these regulations initially going into effect on August 14, 2020 [25], and later receiving a final set of updates on December 10, 2020 [24].

As of this writing, the CCPA has been in force for 14 months, and thus we feel the time is ripe to begin investigating whether data collectors are complying with the CCPA. We argue that assessing compliance with the CCPA is critical for three reasons. *First*, the California Office of the Attorney General (CA OAG) is charged with taking enforcement actions against data collectors who are not in compliance with the law, and researchers can help by identifying cases of non-compliance. *Second*, Californians recently approved a ballot measure known as the California Privacy Rights Act of 2020 (CPRA) [29] that amends and expands on the CCPA, and thus it is important that upcoming rule making around the CPRA be informed by an understanding of what is working, and not working, with respect to the CCPA. *Third*, as of early 2021, there are at least ten other US states considering online privacy legisla-

Maggie Van Nortwick: Northeastern University, E-mail: vannortwick.m@northeastern.edu
Christo Wilson: Northeastern University, E-mail: cbw@ccs.neu.edu

[1] We discuss the criteria that determine whether the CCPA applies to a business in § 2.

[2] Earlier, defunct rule-making around the CCPA by the California Department of Justice also permitted the link to say “Do Not Sell My Info” [21]. This is a distinction we explore in § 4.

tion [62], some of which are in advanced stages of development [58, 67]. It is not too late to update the text of these proposed laws based on insights drawn from the experience of the CCPA.

In this study we take a first step towards these goals by studying the most visible aspect of the CCPA: the DNSMPI hyperlinks that websites covered by the CCPA are required to place on their homepage. We focus on the DNSMPI links because: (1) they are a crucial aspect of the law; (2) they are simple and straightforward to implement in HTML, thus there is no technological burden that prevents website owners from adopting them; (3) they are relatively unambiguous to identify at-scale using automated techniques; (4) we view them as a bellweather for deeper compliance with the law, i.e., we suspect that websites that do not have the DNSMPI link on their homepage are unlikely to be obeying with the CCPAs more complex provisions like data access and deletion rights. We aim to investigate the following research questions:

- **RQ1**: What fraction of major websites now include a DNSMPI link?
- **RQ2**: Of the websites that do not include a DNSMPI link, how many are likely to be non-exempt from the CCPA, i.e., out of compliance?
- **RQ3**: Do websites with a DNSMPI link display it to all visitors, or are websites using geofences to selectively restrict access outside California? If geofences are being used, are they implemented (**RQ3a**) client-side or (**RQ3b**) server-side?
- **RQ4**: Have DNSMPI links changed over time, either in terms of quantity of websites adopting them (**RQ4a**) or phrasing of the DNSMPI links themselves (**RQ4b**)?
- **RQ5**: How are websites choosing to present DNSMPI links, and how does this presentation compare to terms of service (ToS) and privacy policy (PP) links?

To answer these questions, we crawled the homepages of the top 1M domains from the Tranco top list [60] and applied observational analysis methods to identify and examine DNSMPI links. To estimate whether each crawled website was covered or exempt from the CCPA, we rely on two datasets: unique visitor count estimates that we purchased from a marketing analytics firm named Semrush[3], and resource inclusion trees [7, 10, 13] captured by our crawlers that we use to detect the presence of online advertising and/or trackers embedded in each website. These datasets serve as useful, but imperfect, proxies for assessing CCPA applicability—we discuss these limitations in § 6.1. Additionally, our crawls captured website homepages before and after JavaScript rendering, and from IP addresses inside and outside California, to assess the impact of geolocation on the visibility of DNSMPI links. Lastly, by collecting two longitudinal snapshots of website homepages that coincided with key rule-making events by the CA DoJ [24, 25], we can observe changes in CCPA compliance over time.

Overall, we find that there are thousands of popular websites that are likely to be covered by the CCPA (i.e., non-exempt), but have not implemented DNSMPI links. Although we find that DNSMPI adoption is growing over time, the pace of adoption is slow—on the order of 600 websites per month. Furthermore, among websites that have adopted DNSMPI links, we observe many that have implemented user interface designs that may hinder Californians' ability to access their CCPA rights, including difficult to read links and geofences that dynamically hide DNSMPI links. We conclude our study with suggestions for how lawmakers and regulators may encourage compliance with the CCPA (and CPRA) and discourage anti-consumer design practices.

2 Background

We begin by discussing the provisions of the CCPA from the California Civil Code (CA Civ. Code) and the associated rules from the CA DoJ (11 CA ADC). This motivates and informs the design of our study.

2.1 Overview

The CCPA, introduced as CA Assembly Bill No. 375 [27], was passed by the California legislature and signed into law by the CA governor on June 28, 2018. It was subsequently amended by CA Senate Bill No. 1121 [28] on September 23, 2018. The provisions of the law became operative on January 1, 2020.

Broadly speaking, the CCPA grants California residents the right to know who is collecting their "personal information" (PI) and for what purpose, to request the data be deleted, to opt-out of collection, and to not face discrimination for exercising their opt-out right. CA Civ. Code §178.140(o)(1) defines PI as "information

[3] https://semrush.com

that identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household," and provides a non-exclusive list of example data types that qualify, including IP address, browsing history, geolocation, and any inferences drawn from these data.

The CCPA allows Californians to opt-out of the "sale" of their PI, but the definition of "sale" is extremely broad. CA Civ. Code §178.140(t)(1) states:

> "Sell," "selling," "sale," or "sold," means selling, renting, releasing, disclosing, disseminating, making available, transferring, or otherwise communicating orally, in writing, or by electronic or other means, a consumer's personal information by the business to another business or a third party for monetary or other valuable consideration.

This broad definition of "sale" is meant to encompass the routine exchange of personally-identifiable data between websites for monetary value or value consideration, e.g., via techniques like real-time bidding and cookie matching [13]. Unless otherwise specified, when we use the terms "sale" or "sell" in this paper we adopt the stance of the CCPA that encompasses data sharing.

CA Civ. Code §1798.155(b) directs the CA OAG to enforce the provisions of the CCPA. Further, CA Civ. Code §1798.185 directs the CA OAG to develop additional rules and regulations to help businesses comply with the law, and help consumers exercise their rights.

2.2 Applicability

The CCPA applies to any business that collects PI from Californians and meets any of three criteria given in CA Civ Code §1798.140(c)(1):

1. Has annual gross revenue ≥$25M USD,
2. Sells the PI of ≥50K Californians annually, or
3. Derives ≥50% of its annual revenue from the sale of Californians' PI.

Note that nonprofit businesses and government entities are exempt from the CCPA.

In this study, we rely on multiple sources of data to estimate whether websites meet the above three-pronged criteria. We refer to websites that fail to meet these criteria as *exempt* from the CCPA. We introduce these datasets in § 3 and estimate applicability in § 4.

Privacy policies serve several important functions under the CCPA. For example, 11 CA ADC §999.306 (d) [22] states:

> A business does not need to provide a notice of right to opt-out if: (1) It does not sell personal information; and (2) It states in its privacy policy that it does not sell personal information.

In theory, this means that a business may publicly self-certify that it is exempt from the CCPA. Although this privacy policy language seems like it might provide a useful, unambiguous signal for identifying websites that claim to be CCPA exempt, we found it difficult to leverage in practice (see § 6.2.2).

2.3 Requirements and Responsibilities

Businesses that meet the applicability criteria of the CCPA must meet a number of requirements. The most visible requirement, stated in CA Civ. Code §1798.135(a)(1), is that businesses "provide a clear and conspicuous link on [their] Internet homepage, titled "Do Not Sell My Personal Information," to an Internet Web page that enables a consumer [...] to opt-out of the sale of the consumer's personal information."[4][5] The intent of the DNSMPI link is to make it simple for Californians to opt-out of the sale of their PI. The CA DoJ's regulations go further by recommending standard iconography to accent DNSMPI links (11 CA ADC §999.306(f) [24]), and mandating that the opt-out user interface be simple, accessible, and free from *dark patterns* [16, 39] (e.g., 11 CA ADC §999.306(a)(2) and §999.315(d) and (h) [24]). Given their importance, DNSMPI links are the focus of our study.

The CA DoJ's guidance around the DNSMPI link has changed over time. The first two iterations of the regulations permitted the link to say "Do Not Sell My Personal Information" or "Do Not Sell My Info" [23, 25], but the third iteration removed the latter option [26]. This variability over time creates an opportunity for a natural experiment, which we explore in § 4.

Businesses must have a privacy policy on their website to comply with the CCPA, since they are required, per CA Civ. Code §1798.130, to include text reminding Californians of their rights under the CCPA, and to

[4] The CA DoJ also provides guidance for mobile apps to comply with the DNSMPI link requirement [24], but apps are not the focus of our study.

[5] CA Civ. Code §1798.140(l) states ""Homepage" means the introductory page of an Internet Web site". In this study, we assume the root (/) of an HTTP/S domain contains a websites' homepage, or redirects to it.

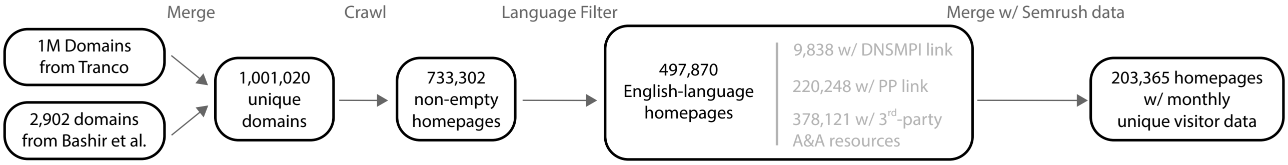

Merge
Crawl
Language Filter
Merge w/ Semrush data
1M Domains from Tranco
2,902 domains from Bashir et al.
1,001,020 unique domains
733,302 non-empty homepages
497,870 English-language homepages
9,838 w/ DNSMPI link
220,248 w/ PP link
378,121 w/ 3rd-party A&A resources
203,365 homepages w/ monthly unique visitor data

Fig. 1. Overview of the data collection steps in our July–August 2020 crawl. Our November–December 2020 data collection re-crawled the 497,870 English-language homepages, while our May 2021 data collection just re-crawled homepages that we had previously identified as containing DNSMPI links.

disclose information about their data collection, selling, and sharing practices. We use the presence of a privacy policy link on a website's homepage as a signal of engagement with online privacy regulations in § 4.

The CCPA lists a number of other requirements for businesses that enable people to exercise their rights to access and delete collected personal information. These facets of the CCPA are not the focus of our study, so we elide further details.

2.4 CPRA

In November 2020, voters in California passed the California Privacy Rights Act (CRPA) as a ballot initiative [29]. The CPRA strengthens the CCPA by closing loopholes in the original text related to data sharing, further regulating the relationships between first- and third-party data collectors, adding additional restrictions on the collection and sale of "sensitive" PI, and creating a California Privacy Protection Agency to enforce the CPRA.

Although the CPRA is not operative until January 1, 2023, its provisions do have two potential implications for our study:

1. The amended §1798.140(d)(1)(B) raises the applicability criteria for businesses from selling 50K Californian's PI to 100K, and
2. The amended §1798.135(a)(1) states that the required opt-out link on website homepages be titled "Do Not Sell or Share My Personal Information".

We explore both of these changes to the law in § 4.

3 Methods

In this study, our goal is to study compliance with the CCPA through the lens of DNSMPI link adoption by websites. To implement our study, we require a variety of data from major websites, as well as some key additional forms of meta-data. In this section, we describe the methods we use to gather these datasets.

3.1 Corpus Selection

We begin by selecting the corpus of websites to crawl for our study. Overall, we aim to focus on websites that are either popular or are known to be part of the online advertising ecosystem. We focus on these two classes of websites because their privacy practices are likely to impact the maximum number of people.

To build our corpus, we joined the top 1 million domains from the research-oriented Tranco[6] domain popularity ranking [60] with 2,902 domains that were identified as third-party trackers and/or advertisers by Bashir et al. [15].[7] To further narrow this list, we performed an initial crawl in which we attempted to resolve each domain to a website, scrape its homepage, extract the page's text, and then analyze the text with the Python `langdetect` library. Figure 1 shows an overview of this process. Our crawler failed to retrieve a non-empty webpage from 267,718 (27%) of the domains in our initial list due to a variety of errors, including DNS resolution failure, connection failures, TLS errors, and HTTP 4XX and 5XX responses. This crawling error rate is expected for a crawl based on the Tranco list, since it contains popular domains that may not host websites (e.g., `windowsupdate.com`). Our final corpus of 497,870 domains includes those that successfully returned an HTML webpage containing English text. We discuss the technical details of our initial and subsequent crawls in detail in § A.1.

[6] We use the Tranco list dated July 17th, 2020, ID 8KVV.
[7] 1,882 tracking domains overlapped the Tranco list, yielding 1,020 unique additions.

3.2 Crawling Web Data

Next, we performed three web crawls to collect a variety of information from websites in our corpus. Each crawl used custom Python scripts to drive Chrome via the Chrome Developer Protocol. After visiting each webpage our crawler waited 15 seconds before attempting to examine the contents of pages (e.g., to detect DNSMPI hyperlinks) to allow page elements to load dynamically and JavaScript to execute.

We conducted our first crawl in July–August 2020 from IP addresses at Northeastern University in Boston.[8] At this time we collected the homepages of our targeted websites before and after rendering JavaScript to infer the presence of client-side logic that selectively shows or hides DNSMPI links (**RQ3a**). Additionally, our crawler parsed the homepages to identify and recursively crawl DNSMPI, ToS, and PP hyperlinks. We configured the crawler to visit at most 20 such hyperlinks. We use the text from these policy pages in § 6.2.2 to briefly examine the language that companies use in their policies to claim CCPA exemption. We describe how we identified relevant hyperlinks in § 3.3.

During this first crawl, we also collected the *inclusion trees* for each homepage [7, 13]. In the inclusion tree for a webpage, the nodes are objects (and their URL origins) that make up the page (e.g., HTML, images, JavaScript, etc.) and edges correspond to causal resource loading relationships. For example, if the root HTML for a webpage uses a `<script src="...">` tag to include a JavaScript file, this would yield two nodes connected by an edge. Similarly, if JavaScript code dynamically inserted an image into a webpage, this would also correspond to two nodes connected by an edge. We use this inclusion tree data to identify websites that include resources from third-party advertising and analytics (A&A) domains, and are thus engaged in data selling behavior that implicates the CCPA (**RQ2**). We describe how we identified A&A domains in § 3.4.

To examine server-side code that selectively shows DNSMPI links (**RQ3b**) and longitudinal changes in CCPA adoption (**RQ4**), we conducted a second crawl in November–December, 2020. This crawl was actually two crawls run in parallel, using IP addresses from Boston and California (rented from Amazon Web Services), respectively. Both browsers were configured to report a California location via JavaScript, but otherwise the two browsers shared no state. We engineered the crawl this way to isolate cases where code on the server-side was using IP address geolocation techniques to identify people in California and serve them pages containing DNSMPI links. If both crawlers were served pages containing a DNSMPI link by a given website, this demonstrated that the web server was not dynamically removing the DNSMPI link for people outside California.

Finally, to examine the presentation of DNSMPI, ToS, and PP hyperlinks (**RQ5**), we conducted a third crawl of homepages from our corpus in May 2021. This crawl only covered the subset of domains on which we had previously observed DNSMPI links. During this crawl we recorded style information from each homepage (after allowing the page 15 seconds to load) including: (1) the overall homepage dimensions; (2) font size, font color, background color, x and y coordinates, width, and height of DNSMPI, ToS, and PP hyperlinks; (3) and font size, font color, and background color of all text on the homepage.

For additional details about our crawlers, see § A.1.

3.3 Identifying Relevant Hyperlinks

To successfully crawl and analyze data for our study, we must be able to identify various specific hyperlinks on webpages. We adopted and generalized the approach used by prior work to identify DNSMPI links and links to privacy policies [4, 52]: at a high-level, on each webpage our crawler extracted the text from each hyperlink and searched it for key phrases. We searched for longer, more specific phrases first, and less specific phrases second, to minimize false positives. After extracting relevant links, we filtered out selected links that contained a list of exclusion phrases as an additional step to reduce false positives. We describe our key and exclusion phrases next.

With respect to identifying DNSMPI links, we searched for eight phrases: “**do not sell my personal information**”, “do not sell my information”, “**do not sell my info**”, “do not sell my personal info”, “**do not sell or share my personal information**”, “do not sell or share my information”, “do not sell or share my info”, and “do not sell or share my personal info”. Only two phrases correspond precisely with the CA DoJ’s regulations, and one corresponds to the text of the CPRA (all highlighted in bold). However, given that the guidance and regulation around the DNSMPI link has changed over time, we decided to take a relatively permissive approach to detecting DNSMPI links. In other words,

[8] Our decision to crawl from Boston IP addresses for this initial crawl leads to limitations that we discuss in § 6.1.

we assume that the operator of a website that includes at least one of these eight DNSMPI links is making a good-faith attempt to comply with the CCPA/CPRA.

With respect to detecting ToS and PP hyperlinks, we provide our list of phrases and exclusion terms in § A.2. We drew these phrases from manual analysis of webpages (the same construction method used by prior work) [4, 52].

Validation. To assess the false positive and negatives rates of our approach for detecting DNSMPI links, we manually examined 250 randomly selected websites from our corpus where we detected a DNSMPI link, and 250 where we did not. We discovered zero false positives, which is to be expected since our list of phrases and exclusion terms was already developed and tuned by hand. We discovered five false negatives: one website that included a DNSMPI link that we did not automatically detect because it was of the form "`do not sell my <a href="...">personal info</a>`" (i.e., the entire phrase was not within the anchor tag), and four websites that included a link with the phrase "Do not sell my data". Based on this manual assessment, we are confident that our DNSMPI link detection approach has high precision and specificity.

3.4 Identifying A&A Domains

The CCPA and CPRA only apply to websites that collect peoples' PI. Although there is no general method for determining whether an arbitrary website is collecting PI, in this study we use the presence of third-party A&A domains as a signal of PI collection. Specifically, we assume that any website that embeds third-party A&A resources is implicated in the sharing and/or selling of PI,[9] and thus must comply with the CCPA/CPRA unless they do not meet to eligibility criteria given in § 2.2.

To identify resource inclusion from third-party A&A domains, we apply two heuristics to the HTTP requests in our inclusion trees. First, we match each requested URL against the EasyList and EasyPrivacy[10] block lists [10, 11, 13]. Second, we apply the methods developed by Fouad et al. to identify inclusions of tracking pixels [35]. If either heuristics "hits" then we label the corresponding website as including A&A resources.

3.5 Estimating Unique Visitors

The most challenging facet of our study is determining which websites are exempt from the CCPA and CPRA. Recalling the three applicability criteria from CA Civ Code §1798.140(c)(1) (see § 2.2), two turn on business revenue. To the best of our knowledge, there is no way to determine the revenue earned by an arbitrary website (especially outside the small set of websites that are obviously owned by public companies, e.g., `google.com` and `facebook.com`). Thus, in this study we focus on the third applicability criteria: the number of Californians whose data is shared or sold by a company.

To estimate the number of unique Californians visiting websites in our corpus, we use data from the marketing research company Semrush.[11] In September 2020 we purchased a "custom report" from Semrush for $750 that contained the unique monthly US visitors to 203,365 of the 497,870 domains in our corpus in August 2020 (Semrush did not have visitor estimates for the remaining domains). These visitor counts are based on Semrush's proprietary measurement and estimation methods. We scaled the unique visitor counts down by 88%, since California accounts for 12% of the internet-enabled US population.[12] Thus, we assume that unique visitors to websites are uniformly distributed throughout the US. We revisit this limitation in § 6.1.

4 Analysis

In this section, we leverage our crawled data to investigate our five research questions.

4.1 Overall DNSMPI Link Adoption

We begin by addressing **RQ1**: what fraction of major websites include a DNSMPI link? Figure 2 presents the fraction of websites in our first crawl (in July–August 2020) that included a DNSMPI link, bucketed into groups of 25K by Tranco rank.[13] For now, we do not distinguish between DNSMPI links that had dynamic visibility via geofencing (for this, see § 4.3) or different

[9] In this work we do not consider "service provider" exemptions to data sharing, see § A.3.

[10] https://forums.lanik.us/

[11] https://www.semrush.com/

[12] https://www.ntia.doc.gov/data/digital-nation-data-explorer

[13] All figures in the paper that group websites by Tranco rank use 25K buckets.

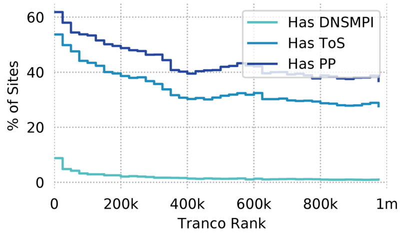


Fig. 2. Percentage of websites in our corpus that include DNSMPI, ToS, and PP links, bucketed by Tranco rank.

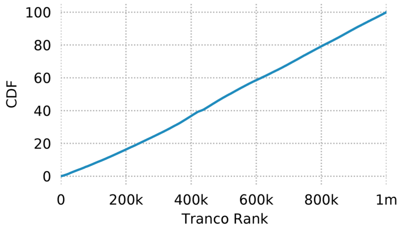


Fig. 3. CDF of websites in our corpus that are missing unique visitor count data from Semrush, sorted by Tranco rank.

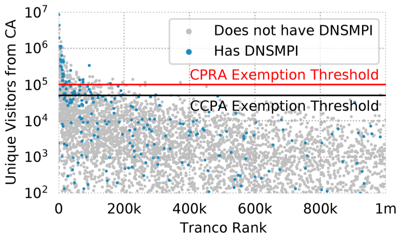


Fig. 4. Scatter plot comparing Tranco rank and estimated unique visitors from California in August 2020 for websites with and without DNSMPI links.

textual content (for this, see § 4.4)—we consider a website as having a DNSMPI link if its homepage HTML contains a hyperlink containing any of the eight phrases given in § 3.3.[14]

In total, 9,838 websites (2%) in our first crawl contain a DNSMPI link. This number is likely an underestimate of the number of top websites that included a DNSMPI link in July–August 2020, due to the impact of IP address-based geofencing of DNSMPI links. As we discuss in § 4.3, 7% of websites in our November–December 2020 crawl dynamically hid their DNSMPI links based on IP address geolocation. Since our July–August dataset was crawled using an IP address in Boston, this suggests that ∼10,500 top websites may have had DNSMPI links in July–August 2020.

As shown in Figure 2, adoption is not uniformly distributed: 9% of the top ranked 25K websites in our corpus adopted DNSMPI links, while the long tail of websites levels out at ∼1% adoption. We hypothesize that several factors drive these trends: (1) popular websites may be more likely to be actively maintained, and are thus more responsive to changes in the law; (2) popular websites may be more likely to have developer and legal resources to implement privacy-compliance regimes; and (3) the CCPA may be more likely to apply to popular websites. This last point is crucial: although we find that the level of DNSMPI link adoption is quite low overall, this does not necessarily imply that the vast majority of websites in our corpus are out of compliance with the CCPA. We examine CCPA applicability in § 4.2.

In comparison to DNSMPI links, 220,248 (44%) websites in our corpus contain a PP link, and 172,039 (35%) contain a ToS link.[15] We hypothesize that the relatively low adoption of DNSMPI links versus PP and ToS links is partially explained by the newness of the CCPA—in contrast, the US FTC has been urging websites to include privacy policies since the 1990's [73]. 96% of websites in our corpus with a DNSMPI link also contain a PP link, which is crucial since the CCPA requires that compliant websites include both things.

Recall that we include 2,902 domains in our corpus that belong to third-party A&A companies (see § 3.1). Given the nature of these companies' business model, the CCPA almost certainly applies to them. 149 of these websites (9%) contained DNSMPI links, which is similar to the adoption rate of the top 25K websites overall. It is disheartening to see so few A&A companies complying with the DNSMPI link requirement of the CCPA.

4.2 DNSMPI Adoption vs. CCPA Exemption

In § 4.1 we observe that DNSMPI link adoption is very low overall. This motivates **RQ2**: of the websites that do not include a DNSMPI link, how many are likely to be non-exempt from the CCPA, i.e., out of compliance? Determining whether the CCPA applies to a given website is challenging given that the data required to make a determination is rarely publicly available (see § 2.2). Thus, we take two approaches to assessing CCPA applicability by examining (1) unique visitors to websites and (2) inclusion of resources from A&A domains. In this section, we continue to leverage data from our July–August web crawl.

[14] We found only one website, https://swlaw.com, containing the CPRA link phrasing, possibly because it is a law firm.

[15] Other studies have identified broadly similar percentages of PP links on websites. Linden et al. identified PP links on 29% of top European websites and 64% top websites globally [53]. Amos et al. identified PP links on 55% of top ranked, English-language websites drawn from the Alexa ranking [4].

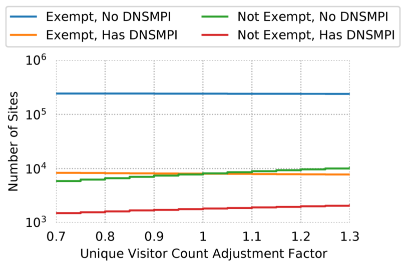


Fig. 5. Effect of adjusting the estimated unique visitors from California on the number of websites that are CCPA exempt.

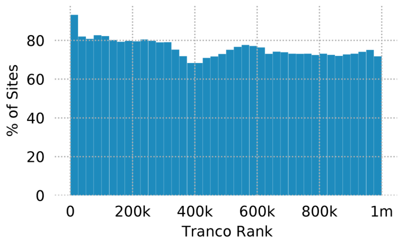
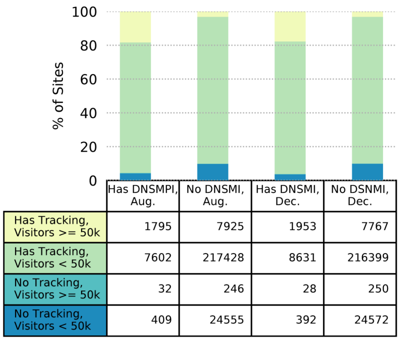


Fig. 6. Percentage of websites in our corpus that embed third-party A&A resources, bucketed by Tranco rank.



	Has DNSMPI, Aug.	No DNSMI, Aug.	Has DNSMI, Dec.	No DSNMI, Dec.
Has Tracking, Visitors >= 50k	1795	7925	1953	7767
Has Tracking, Visitors < 50k	7602	217428	8631	216399
No Tracking, Visitors >= 50k	32	246	28	250
No Tracking, Visitors < 50k	409	24555	392	24572

Fig. 7. The percentage and count of websites in four dimensions: (1) inclusion of a DNSMPI link, (2) inclusion of 3^{rd}-party A&A, (3) having ≥50k unique visitors from CA, and (4) time.

4.2.1 Unique Visitors

CA Civ Code §1798.140(c)(1) states that the CCPA only applies to companies that share or sell the PI of ≥50K Californians per year (the CPRA ups this threshold to ≥100K). In this section, we estimate which websites in our corpus these rules may apply to by using unique visitor count data that we purchased from Semrush (see § 3.5).

Recall that Semrush only had unique visitor data for 203,365 (41%) of the websites in our corpus. Figure 3 plots the CDF of websites with missing visitor data, sorted by Tranco rank. In absolute terms, 84% of the websites with missing data have Tranco rank ≥200K, which, as we show in Figure 4, means they are unlikely to be popular enough to be covered by the CCPA. In the remainder of this section, we focus on the 203,365 websites for which we have Semrush data.

To investigating the relationship between Tranco rank, unique visitors from California, and adoption of DNSMPI links we plot Figure 4. We make three observations from this figure. *First*, we observe that there is a statistically significant correlation between unique visitors counts and Tranco rank (Pearson's $r = -0.042^{***}$).[16] This is expected, given that the Tranco ranking includes data from underlying sources like Alexa,[17] which are themselves compiled from web browsing history data. We also observe that the correlation between unique visitors and Tranco rank is stronger for websites that include DNSMPI links ($r = -0.16^{***}$) than those without ($r = -0.033^{***}$). These statistical results lend face credibility to the unique visitor data from Semrush and further reinforce the relationship between website popularity and DNSMPI link adoption.

Second, we can see in Figure 4 that 95% of websites in our corpus fall below the CCPA and CPRA exemption thresholds. This helps explain why we find that DNSMPI link adoption is so low overall. Of the websites in our corpus that did adopt a DNSMPI link, 1,827 (19%) are above the CCPA threshold, while 1,238 (13%) are above the CPRA threshold. On one hand, these results suggest that the CCPA is "working", in the sense that popular websites to which it likely applies have been adopting DNSMPI links. On the other hand, these results also highlight how far we have yet to go to reach full adoption of the CCPA: of the 9,998 websites above the CCPA threshold, only 18% have adopted DNSMPI links. Similarly, of the 5,417 websites above the CPRA threshold, only 23% have adopted DNSMPI links.

Third, we note that doubling the exemption threshold from 50K to 100K cuts the number of websites that the privacy law is likely to be applicable to by 54%. This has clear, and we argue negative, implications for Californian's privacy rights.

Simulations. A major limitation of our CCPA applicability analysis thus far is that our estimates of unique visitors from California may be incorrect for a number of reasons, including: (1) Semrush's data may have errors; (2) the Semrush data we purchased only covers visitors in August 2020, so unique visitor counts for the full year may be higher; and (3) we assume that unique visitors are uniformly distributed throughout the US, but a given website may be more or less popular with Californians. To address these limitations, we per-

[16] $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
[17] https://www.alexa.com/topsites

form simulations where we adjust the estimated number of unique visitors from California by fractions in the range $[0.7, 1.3]$ and re-evaluate the number of websites that the CCPA would and would not be applicable to.

Figure 5 shows the results of our simulations. As expected, as the adjustment factor grows (i.e., the number of unique visitors from California pwe website grows) the number of websites that are not exempt from the CCPA grows as well. In our corpus the number of websites that are not exempt and include a DNSMPI link grows from 1,499 to 2,095, while non-exempt websites without DNSMPI links grows from 5,886 to 10,436. Conversely, the number of exempt websites with DNSMPI links only shrinks from 8,339 to 7,743, i.e., the vast majority of websites that flip from exempt to non-exempt do not include a DNSMPI link.

Overall, this simulation demonstrates that the number of websites that we suspect to be out of compliance with the CCPA's DNSMPI link requirement is highly sensitive to our estimates of unique visitors. This represents a serious challenge that stems from the formulation of the CCPA (and CPRA): it is hard to know which websites the law applies to. This is definitely true for outsiders like academics and regulators, and it may even be challenging for the websites themselves to know. We revisit this challenge in § 6.2.2.

4.2.2 Inclusion of A&A

Although we have shown that many websites with ≥50K unique visitors from California do not contain a DNSMPI link, it may be that these websites are still exempt from the CCPA (and CPRA) because they do not sell or share Californians' PI. There is no way for us to say, definitively, whether any given website sells Californians' data, but we can use the presence of third-party A&A domains as a proxy. In other words, we assume that any website that embeds resources from a third-party A&A domain is likely to be selling data. We argue that if a website includes A&A resources and has ≥50K unique visitors from California, this provides compelling evidence that the CCPA applies to it. In this analysis we do not consider "service provider" exemptions for data sharing, a limitation that we discussion in § 6.1.

Figure 6 shows the percentage of websites in our first crawl that contained third-party A&A resources (see § 3.4 for our A&A detection approach). Overall, 76% of the websites in our corpus embed at least one A&A resource, which is unsurprising given the ubiquity of third-party tracking and advertising observed by prior studies [15, 20, 30–32, 51]. We see that the most highly ranked websites are more likely to include A&A resources, which is possibly due to the increased pressure on popular websites to monetize.

Now that we have established that A&A resource inclusion is endemic in our corpus, we present Figure 7,[18] which is a stacked bar graph that stratifies our corpus in four dimensions: including or not including a DNSMPI link, embedding or not embedding third-party A&A resources, by unique Californian visitors ≥50K or <50K, and by time (either our July–August or November–December crawl).

We highlight five observations from Figure 7. *First*, Figure 7 reconfirms that the vast majority of websites in our corpus embed third-party A&A resources. The percentage of websites with ≥50K unique visitors from California and no A&A resources is so small that it is invisible in the bar graph, i.e., popularity goes hand-in-hand with advertising and tracking. *Second*, we find that the majority of websites in our corpus that include a DNSMPI link do not appear to fall within the law's applicability criteria: they have <50K unique visitors from California, and in some cases also do not embed third-party A&A resources. It is unclear why these websites would expend the effort to comply with the CCPA.

Third, the largest groups of websites by far are those with no DNSMPI links, <50K unique visitors from California, and A&A resource embeds. Despite potentially selling Californians' data, these websites are likely to be exempt from the CCPA and CPRA unless they (1) have annual gross revenue ≥\$25M USD or (2) derive ≥50% of their annual revenue from the sale of Californians' PI. Neither of these stipulations seems likely to apply given that these are relatively unpopular websites. However, A&A websites may be an exception: as we note in § 3.1, 35% of A&A domains on our dataset did not appear in the Tranco ranking, probably because they do not receive enough first-party visits to appear popular under standard audience measurement techniques. That said, A&A domains almost certainly sell data from ≥50K unique Californians in their capacity as third-parties.

Fourth, the results in Figure 7 lend additional credence to our observations from Figure 4—∼80% of websites in our corpus that have ≥50K unique visitors from California include A&A resources but not a DNSMPI link, suggesting that they are not in compliance with

[18] Figure 7 excludes 1,218 websites that included DNSMPI links in our second crawl but did not have corresponding Semrush visitor data.

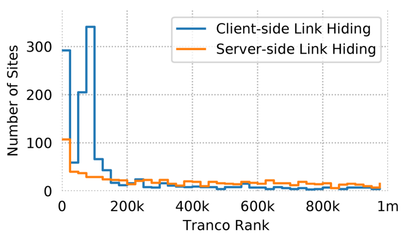


Fig. 8. Number of websites that we observe using client-side, server-side, or both methods to hide DNSMPI links, bucketed by Tranco Rank.

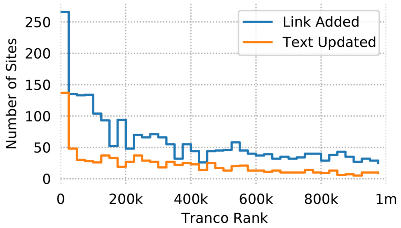


Fig. 9. Number of websites in our corpus that added or updated their DNSMPI link between August and December 2020, bucketed by Tranco rank.

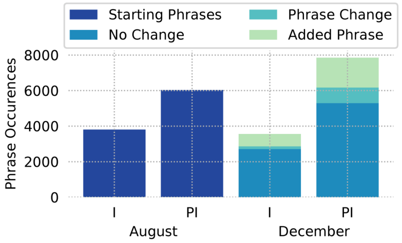


Fig. 10. Number of websites that changed the DNSMPI link phrasing on their homepage over time.

the CCPA. *Fifth*, comparing our data from the first and second crawls we see that the trends are essentially unchanged. We delve deeper into temporal changes in DNSMPI link adoption in § 4.4.

4.3 Dynamic DNSMPI Link Hiding

In this section we examine the dynamic behavior of websites in our corpus by posing **RQ3**: do websites with a DNSMPI link display it to all visitors, or are websites using geofences to selectively restrict access outside California? If geofences are being used, are they implemented (**RQ3a**) client-side or (**RQ3b**) server-side? The broad question of whether geofences are being employed in general has implications for when and how Californians may exercise their rights under the CCPA, which we unpack in § 6.2.3. The specific questions of how these geofences are implemented have implications for the design of any technology that aims to programmatically interact with DNSMPI links, whether it is to study them as we are, or to build user-facing tools around them (e.g., privacy-enhancing browser extensions).

For this analysis we leverage the data from our second crawl, since it included dynamic JavaScript snapshots as well as IP addresses within and outside California. Figure 8 shows the number of websites in our second crawl that used client-side or server-side methods to dynamically hide DNSMPI links for people who appeared to be outside California, bucketed by Tranco rank. Of the 12,222 websites in the second crawl that include a DNSMPI link, 2,101 (17%) exhibit dynamic link visibility. We find that 1,293 (62%) of the websites who adopt link geofencing use client-side methods, while 808 (38%) use server-side methods. Further, we observe that almost all link-hiding practitioners are from the top ranked 100K websites. This finding is heavily dependent on our earlier observation that most DNSMPI link adoption is by top ranked websites. Additionally, we hypothesize that only popular websites may have the resources to implement link-hiding logic.

While we observe relatively few websites adopting dynamic DNSMPI link-hiding behind geofences, we argue that this practice is still problematic. We discuss this issue further in § 6.2.3.

4.4 DNSMPI Links Over Time

In this section we turn to **RQ4**: have DNSMPI links changed over time, either in terms of quantity of websites adopting them (**RQ4a**) or phrasing of the DNSMPI links themselves (**RQ4b**)?

Starting with **RQ4a**, Figure 9 plots the number of websites that we observed adding a DNSMPI link between our first and second crawls (which were roughly four months apart). In absolute terms, we only observe 2,369 websites that added a DNSMPI link, with the vast majority of adoption activity confined to the top ranked 200K websites.[19] During this time period, 793 websites removed their DNSMPI link. Thus, while we can say that DNSMPI link adoption continues to grow, it is doing so at a slow pace.

Turning to **RQ4b**, as we noted in § 2.3, the CA DoJ has changed their guidance around the required text of the DNSMPI link over time—at the start of 2020 two phrasings were allowed [23, 25], but in October 2020 the CA DoJ eliminated one of the two phrasings (the “Do Not Sell My Info” phrasing) [26]. Since the release of the updated guidance falls directly between our two

[19] This count of websites that added a DNSMPI link excludes the 808 sites that did not have the link during our July—August 2020 crawl and then were found to have server-side link-hiding during our November—December 2020 crawl.

crawls, this creates an opportunity for a natural experiment: how many websites are sufficiently engaged with the changing CCPA regulations that they updated the phrasing of their DNSMPI link from "Info" to "Personal Information" (or vice versa)?

Figure 9 shows the number of websites in our corpus that updated the text of their DNSMPI link between our first and second crawls, bucketed by Tranco rank. Out of the 9,838 websites in our corpus that included a DNSMPI link in our first crawl, 3,806 used the "Info/Information" phrasing, and 876 updated their link text by our second crawl. In other words, most websites in our corpus do not appear to be closely tracking the CA DoJ's CCPA guidance over time—rather, it is likely that they updated their website to comply with the CCPA at one point in time and considered that project to be complete. Additionally, we also see that the vast majority of phrasing updates occurred on high ranking websites, which is unsurprising given their higher rates of DNSMPI adoption overall and their (presumably) greater access to development and legal resources.

Figure 10 delves deeper into DNSMPI link adoption and phrasing changes over time by (1) separating websites based on the phrasing of their DNSMPI link (**I** being "Info" and **PI** being "Personal Information") and (2) identifying three actions websites could have taken between our first and second crawls. From the August data, when both phrases were permissible under the CA DoJ's guidance, we see that the "Personal Information" phrasing was already more popular than the "Info" phrasing. By December, 689 websites that previously did not have a DNSMPI link added one with the "Info" phrasing, and another 156 websites switched from the "Personal Information" to "Info" phrasing. Note that these changes may have occurred before the CA DoJ announced their altered guidance in October 2020, so these changes may have made sense at the time. In contrast, by December, 1,680 websites that previously did not have a DNSMPI link added one with the "Personal Information" phrasing, and another 876 websites switched from the "Info" to "Personal Information" phrasing.

Based on Figure 10, we draw two conclusions. *First*, there does appear to be a small number of websites that are adapting their CCPA-compliance posture over time, as evidenced by their switch from an "Info" to "Personal Information" DNSMPI link phrasing. While we cannot rule out that these switches are coincidences that are unrelated to the changing guidance, we feel that this is unlikely given that web designers probably prefer the shorter "Info" link phrasing. However, *second*, although the rate of "Personal Information" phrase adoption appears to be outpacing "Info" adoption, in December 2020 31% of websites in our corpus that had a DNSMPI link still used the outdated "Info" phrasing. This finding does not bode well for the future adoption of DNSMPI links with the CPRA mandated phrasing.

4.5 DNSMPI Link Presentation

Finally, we turn to **RQ5**: how are websites choosing to present DNSMPI links, and how does this presentation compare to terms of service (ToS) and privacy policy (PP) links? CA Civ. Code §1798.135(a)(1) states that DNSMPI links must be "clear and conspicuous" on homepages, but neither the statute nor the CA DoJ's guidance define these two terms precisely. Rather than evaluating the presentation of DNSMPI links to a definitive standard, we instead compare them to (1) the presentation of other text on homepages, and (2) the presentation of ToS and PP links. We structure our analysis based on the US FTC's guidance on clear and conspicuous disclosures, abbreviated using the 4Ps mnemonic [34], of which two points are salient to DNSMPI links: *prominence*, i.e., is text big enough and high enough contrast for the average person to read, and *placement*, i.e., where is text displayed and would a reasonable person be able to find it? For the following analysis we leverage data from our third crawl, which was conducted in May 2021.

Prominence. Figure 11 delves into font sizing by presenting a CDF of font size ratios, computed by dividing the font size for a given policy link (DNSMPI, ToS, or PP) by the most common font size on the corresponding page.[20] Thus, ratios <1 are cases where the policy link has a smaller font size than the majority of text on the corresponding webpage. We find that all three policy links have nearly identical distributions, with the median ratio being ∼0.9. Further, we see that ∼57% of pages have a ratio <1, meaning that the policy links tend to be rendered in a smaller font than other text on the corresponding webpage. The mean font size for policy links in our dataset is 13 points, whereas the mean most common font size is 15 point.

[20] To determine the most common font size for a given webpage, we counted the number of characters of visible text that were rendered in each font size on the page, then selected the font size with the highest count.

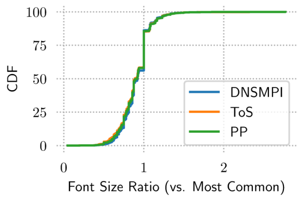


Fig. 11. Ratio of the font sizes for DNSMPI, ToS, and PP hyperlinks versus the most common font size on the corresponding webpages.

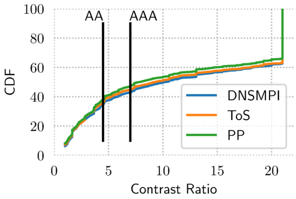
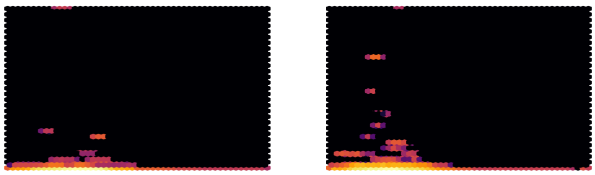


Fig. 12. Contrast ratio between the font and background colors for DNSMPI, ToS, and PP hyperlinks.



(a) DNSMPI

(b) Privacy Policy

Fig. 13. Heat map showing the location of pixels corresponding to DNSMPI and PP hyperlinks in our corpus. All coordinates are normalized to a 528×288 pixel grid (the same aspect ratio as a 1080P display). Yellow is hottest, and the color scale is log_{10}.

The US FTC and the W3C Web Accessibility Initiative (WAI)[21] do not specify a minimum font size that is necessary for text to be clear, conspicuous, or accessible. Instead, the WIA does specify minimum contrast ratios between font and background colors. Specifically, the Web Content Accessibility Guidelines (WCAG) 2.2 define three contrast ratio ranges [1]:

- $< 4.5:1$ is not accessible,
- $[4.5:1, 7:1)$ is accessible with grade AA, and
- $\geq 7:1$ is accessible with grade AAA.

In Figure 12 we plot the CDFs of contrast ratios for our three types of policy links. Once again, all three links exhibit almost the same distributions overall. ∼40% of the policy links in our dataset have contrast ratios below AA level, which may not meet the bar for being clear and conspicuous. Another 5–10% reach the AA level, while the majority reach the AAA level. ∼40% of the policy links in our dataset have a ratio of 21:1 (the maximum), denoting black-on-white or white-on-black text.

Placement. In Figure 13 we investigate the placement of DNSMPI and PP links on homepages collected during our third-crawl.[22] Each subfigure is a heat map showing the most frequent pixel coordinates that are occupied by the respective policy links. To compute these subfigures, we took each homepage and normalized its width and height into a 528×288 pixel space, which we chose because it maintains the aspect ratio of a typical 1080P display. Next, for each homepage, we computed the pixels covered by each policy link in the normalized space, and present the final counts in log_{10} space in Figure 13. Note that our normalization procedure has less distortion in the x dimension than the y, because most webpages do not have content that scrolls horizontally (i.e., the content is the width of a typical computer monitor), but most webpages do scroll vertically.

We observe that the general placement patterns for the DNSMPI and PP links are similar, and that roughly speaking they tend to fall in one of three places: in the top-of-page navigation bar, in the page body, or in the bottom-of-page footer. The footer option is most common by orders of magnitude. The homepages that have policy links within the page body represent cases where policy links were present within a pop-up "consent banner"-style modal dialog.

Overall, our placement findings paint a complicated picture. On one hand, page footers are not a prominent location, which may discourage Californians from exercising their CCPA rights. On the other hand, placing ToS and PP links in the page footer is such a common user interface design pattern that perhaps Californians who are interested in exercising their CCPA rights will intuitively know to check the page footer for a DNSMPI link. Further study is necessary to determine the effect of DNSMPI link placement on Californians' ability to exercise their CCPA rights.

5 Related Work

In this section, we survey related work at the intersection of data collection, online privacy, and regulation.

Privacy Policies. Since the US Federal Trade Commission adopted the so-called "fair information practice principles" [73] in the late 1990's, *privacy policies* have become the de-facto standard for data collectors to communicate their practices to people. Studies have repeatedly found that privacy policies are widely adopted by

[21] The W3 WAI is the primary standards body that ratifies accessibility-related standards for the web.

[22] We omit results for ToS links because they are nearly identical to the DNSMPI and PP link placement results.

online services [4, 31, 57, 66, 68]. However, until recently the form and content of privacy policies was effectively unregulated, and thus researchers have documented a wide range of problematic practices, including: interface designs that make privacy policies difficult for people to find [48]; the use of interface designs that decontextualize privacy choices [2]; and policy language that is intentionally long, legalistic, and unrealistically positive [33, 48, 61].

The CCPA, like the GDPR, includes a number of requirements for the information that data collectors must disclose in their privacy policies, as well as guidelines about the form of these disclosures. Linden et al. examined snapshots of over 6,000 privacy policies and concluded that the GDPR had an overall positive effect on privacy policies, with data collectors making more disclosures, more specific disclosures, and adopting language that was more understandable [53]. Tang et al. found that, despite mandates that companies make privacy policies more understandable, the use of technical jargon in policies is still widespread and is confusing for people [69]. It remains unclear the extent to which data collectors are complying with the privacy policy mandates of the CCPA.

Given that privacy policies are rich source of information about privacy practices, there have been numerous efforts to leverage this data. Researchers and activists have used techniques like crowdsourced annotation [75, 76] and machine learning to extract structured knowledge from privacy policies. This structured data can then be used to build tools that enhance peoples' understanding of privacy policies [18, 43, 70, 77], analyze policies for internal inconsistencies [5], and automate the discovery of policy violations by comparing declared data practices with network traces of data flows [6, 52, 78, 79].

Cookie and Tracking Consent Notices. The EU e-Privacy Directive and the GDPR require that data collectors receive explicit consent from people before setting cookies in their user agents or collecting personal data. Numerous studies have catalogued the proliferation of "cookie bars" and pop-up *consent notices* across websites and apps in response to these laws [68]. For instance, Degeling et al. conjecture that the GDPR caused the fraction of top websites that had adopted a consent notice to increase from 46% to 62% [31]. These adoption rates are significantly higher than the adoption rates for DNSMPI links that we observe in § 4. Additionally, studies have found that these regulations instigated technical changes with respect to when, how many, and who sets cookies in user agents [30, 31, 65]. However, despite these laws, researchers have observed widespread non-compliance [56]—one study observed 49% of websites in their sample setting cookies before a person has given affirmative consent [71].

Although these consent notices are supposed to empower people to exercise control over data collection, numerous studies have documented services using confusing and deceptive *dark patterns* [19, 39] in their consent notices [59, 63, 74]. Experiments have confirmed that these designs are effective at increasing the likelihood of people consenting to data collection [9, 37, 74]. Gray et al. present an extensive discussion of the design ethics and legality of GDPR consent notices in [40].

Data Access, Deletion, and Opt-out. The rights granted to Californians under the CCPA are similar to those granted to European citizens under the GDPR, including the right to access and delete data that has been collected about them. Unfortunately, researchers have identified problems with the processes that data collectors have adopted to implement these rights, including: interface designs that are hard for people to locate and use effectively [3, 42]; data processors that ignore or are slow to respond to data access requests [72]; and systems that rely on weak authentication methods and thus leak sensitive data to unauthorized parties [17, 55].

The right to opt-out of data collection is also codified in the GDPR and CCPA. Researchers have begun developing systems that identify and surface these opt-out choices to make them more accessible to people [49, 64]. These systems are necessary, at least in part, because data collectors have adopted interface designs that make it difficult for people to exercise their opt-out rights [36, 41, 42].

To the best of our knowledge, there are not yet studies that look specifically at data collectors' compliance with the data access, deletion, and opt-out mandates of the CCPA at-scale. Given the lack of compliance with the DNSMPI link requirement that we observe in this study, we hypothesize that many data collectors are not yet complying with these mandates either. This suggests a need for future, follow-up work to assess whether these requirements are being met, and whether the mechanisms and interface designs being used to effect these rights are secure, usable, timely, functionally correct, and obey the guidelines stipulated in the law (e.g., CA Civ. Code §1798.130 [28] and 11 CA ADC §999.306 [22]).

6 Discussion

In this study, we collected large-scale crawled data and analyzed it to investigate various aspects of CCPA compliance that all center around DNSMPI links. To summarize our key findings:

- **RQ1**: Of the 497,870 reachable, English-language websites in our corpus, we find that only 9,838 (2%) adopted a DNSMPI link, with highly popular websites being far more likely to adopt.
- **RQ2**: Once we account for the CCPA's applicability requirements, we find that there are only 7,767 websites in our corpus (in December 2020, out of 203,365 for which we have visitor data) that (1) have ≥50K unique visitors from California and (2) include third-party A&A resources. Using simulations to account for limitations of our dataset, we demonstrate that the number of websites that may be out of compliance with DNSMPI link requirement may vary from a few thousand to just over ten thousand. We estimate that moving to the CPRA's 100K unique visitor criteria will cut the number of websites that are covered by the law by ∼50%.
- **RQ3**: We identify 2,101 websites that are using geofences to hide their DNSMPI link from people who appear to be outside California, with 62% of these websites adopting a client-side geofence. We observe that popular websites are more likely to adopt geofences.
- **RQ4**: Over the course of four months, the number of websites we observe adopting DNSMPI links increased by 2,369, which suggests that CCPA compliance is growing slowly. Furthermore, we only observe 876 websites that adjusted their DNSMPI link in response to updated guidelines published by the CA DoJ in October 2020. This suggests that many companies are not staying current with changes to US privacy laws.
- **RQ5**: ∼40% of the DNSMPI links in our corpus fail to meet minimum standards for readability, which suggests they may not meet the CCPA's "clear and conspicuous" mandate. We also observe that the vast majority of DNSMPI links are placed in the footer of websites, just like ToS and PP links.

Our crawled datasets and source code for generating the figures that appear in this paper are available here.

In the remainder of this section we discuss limitations of our work, and make recommendations for policymakers based on our findings.

6.1 Limitations

Our estimates of the number of websites in our crawl that are covered by the CCPA are neither upper nor lower bounds. On one hand, the number of covered websites could be higher: there may be websites whose unique visitor counts are under-counted in the Semrush data, or there may be websites that have few visitors but have annual gross revenue over $25M. On the other hand, the number of covered websites could be lower: the Semrush data may over-count unique visitors to websites, and some popular websites may be exempt because they are nonprofits, government agencies, or do not sell data. Additionally, the Semrush data that we rely on may have a selection bias, since websites that are privacy-preserving are unlikely to share visitor data with a marketing firm like Semrush. Although Semrush's data collection practices are a black box, they do provide website usage data for high-profile services like Crunchbase, which provides some face validity to Semrush's data.

We exclude non-English websites from our analysis because we cannot reasonably search for the DNSMPI phrase in an unbounded number of languages. However, the CCPA applies to all websites doing business in California, regardless of the domicile of their business or the language of their website. Future work should examine differences in CCPA compliance stratified by the national origin of websites, similar to work on GDPR compliance across international boundaries [30, 65].

Our decision to conduct our July–August 2020 crawl from IP addresses in Boston was based on the assumption that dynamic DNSMPI link-hiding would be implemented primarily using client-side techniques, and thus not impact our data collection. As we show in § 4.3, this is not the case: 38% of websites in our sample that adopted a dynamic DNSMPI link used server-side methods to implement it. Based on this knowledge, we adjust our estimates of overall DNSMPI link adoption in § 4.1. Furthermore, this highlights the importance of the distinction between **RQ3a** and **RQ3b**, and reinforces the importance of IP address selection for future studies of CCPA and CPRA compliance.

As with any study that relies on automated web crawlers, it is possible that websites detected our crawler and changed their behavior in response, e.g., by adding or removing a DNSMPI link. We used standard precautions to frustrate attempts to detect our crawler, including: using a legitimate User-Agent string; avoiding repetitive, high-rate HTTP requests; executing JavaScript and downloading all sub-resources from webpages; and avoiding well-known, detectable automation frameworks like Selenium.

In this work we only examined compliance with the DNSMPI hyperlink requirement of the CCPA—we did not examine compliance with other facets of the law such as data collection opt-out requests, data access requests, or data deletion requests. That said, during our crawls we collected the webpages that DNSMPI links pointed to, as well as the privacy policies of each crawled website. In future work we plan to explore these data sources to investigate compliance with these additional facets of the CCPA.

For the purposes of this study, we focused on the second CCPA eligibility criteria (sale of PI from ≥50K Californians) from CA Civ Code §1798.140(c)(1). There may be more websites that are CCPA-eligible than we estimate due to meeting the first or third criteria. We considered operationalizing the first criteria (gross revenue ≥$25M USD) for this study but found that this was impractical: although revenue information for *companies* is available from services like Crunchbase and Owler, mapping companies to *domains* they control is an open problem, especially at the scale of our study. Thus, we leave the challenge of assessing CCPA eligibility under the first and third criteria as future work.

In this work we assume that all resource inclusions from third-party A&A domains represent data sharing under the CCPA, but this may overestimate the actual prevalence of such data sharing. As we discuss in § A.3, the CCPA and CPRA provide exemptions for data sharing with third-party "service providers". A&A companies offer products that do and do not fall under this exemption, making it difficult to determine whether all inclusions from A&A domains implicate the CCPA. Future measurement studies of CCPA compliance may be able to decode "service provider" APIs and thus more accurately determine whether communications with third-parties fall under this exemption [56].

6.2 Recommendations for Policymakers

In this section, we synthesize the observations from our study into recommendations for policymakers.

6.2.1 Encouraging Compliance

We wholeheartedly encourage the California Attorney General's Office to begin taking enforcement actions under the CCPA. We hypothesize that compliance rates with the CCPA will improve dramatically once examples have been made of several (prominent) websites. We note that enforcement actions by regulators are particularly crucial under laws like the CCPA and CPRA that do not have broad private rights of action. We hope that our study provides general methods and specific data that are useful for regulators to assess CCPA and CPRA compliance at-scale, and thus inform future enforcement actions.

Another potential avenue for regulators to encourage compliance with the CCPA is through pro-active outreach to website owners (e.g., via contact information in WHOIS), major web hosting services and ad tech firms, and systematically important platform providers. For example, web infrastructure providers like Wordpress, Squarespace, and Wix have direct lines of communication with their customer-bases. Similarly, ad tech firms that provide supply-side platforms and ad exchanges can communicate with, and even demand unilateral changes from [12, 44], their publisher partners. Regulators could encourage these companies to contact their customers and remind them to comply with relevant regulations. Alternatively, many website administrators use tools like the Google Search Console to optimize their website for indexing by search engines. Providers of these tools, like Google, could add a notice inside the tools encouraging their users to comply with relevant regulations. Google in particular has encouraged website administrators to adopt a variety of pro-consumer web technologies in the past, such as mobile device-friendly layouts and HTTPS encryption, by boosting the ranks of websites with those features in Google Search results [8, 54]. Regulators could encourage Google, and other search engines, to adopt similar incentives for websites that comply with the basic tenets of privacy regulations, such as including DNSMPI links on their homepages.

6.2.2 Determining Exemption

As our study demonstrates, determining which websites are covered or exempt from the CCPA is very challenging, which we hypothesize may have negative implications for enforcement of the law. California residents are allowed to file complaints about suspected CCPA violations with the California Attorney General, but it is unclear how an ordinary person would know whether a website is covered by the CCPA or not. Even the California Attorney General may need to seek revenue and visitation data from companies to determine their exemption status before launching into investigations of potential CCPA rules violations.

The CA DoJ's regulations for the CCPA state the following in 11 CA ADC §999.306 (d) [22]:

> A business does not need to provide a notice of right to opt-out if: (1) It does not sell personal information; and (2) It states in its privacy policy that it does not sell personal information.

Thus, in theory, privacy policies may offer guidance on whether a website's owner believes it is exempt from the CCPA. We manually examined dozens of privacy policies to determine whether we could use these declarations to augment our analysis. Although we found that the use of phrases like "we do not sell personal information" is widespread in privacy policies, we also found that these statements are largely useless as guideposts for CCPA compliance (or lack thereof). *First*, privacy policies have used these phrases since before the CCPA turned them into terms of art—as such, we observed many pre-CCPA privacy policies that now read as if they are declaring a website as exempt from the CCPA, when in fact they are not making such a claim. *Second*, the CCPA's definition of the term "sell" remains contentious: we observed post-CCPA privacy policies that use the phrase "do not sell" literally, i.e., the direct exchange of data for money, rather than according to the CCPA's broader definition that includes data sharing. These privacy policies often contain additional declarations about data collection and sharing with third-parties that clearly implicate the CCPA. *Third*, we observed many privacy policies that combine the phrase "do not sell" with a variety of caveats, qualifiers, and exceptions that involve the collection and sharing of data with third-parties. For all of these reasons, we believe that the presence of phrases like "we do not sell" in privacy policies cannot currently be used as a simple litmus test for self-declarations of CCPA exemption.

We encourage lawmakers drafting future CCPA-like laws, amending existing laws, or developing guidance around existing laws to consider mandating that websites self-identify as covered or exempt from the relevant regulation and explain why in their privacy policies using specific and unambiguous language. This would facilitate transparency and accountability by enabling research-at-scale like ours, as well as providing a useful signal for individual people and enforcement agencies looking to file complaints under relevant laws.

6.2.3 Dynamic DNSMPI Links

CA Civ. Code §1798.135(b) permits businesses to maintain separate homepages for Californians and non-Californians, with the former containing the DNSMPI link, so long as "the business takes reasonable steps to ensure that California consumers are directed to the homepage for California consumers" [28]. We observe that hundreds of websites take advantage of this provision in the law.

We believe that future CCPA-like laws should include language banning the practice of geofencing consumer rights notifications like DNSMPI links. The issue of geofencing may eventually become moot if a sufficient number of US states adopt CCPA-like laws and website operators simply give up on the practice, or if the US Congress passes a nationwide online privacy law. But until these things happen (if they happen), lawmakers should consider proactively protecting the rights of their constituents. IP-based geolocation inference is notoriously inaccurate [45, 46], and even if the data is accurate it will still misclassify California residents who (1) happen to be abroad, or (2) adopt privacy enhancing technologies like VPNs or Tor. Similarly, demanding that people enable fine-grained location access for a website via JavaScript as a precondition for accessing privacy options is unreasonable and counter-productive.

7 Acknowledgments

We thank the anonymous reviewers and our shepherd, David Mohaisen, for their helpful comments. This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Accessibility Guidelines Working Group. Understanding success criterion 1.4.3: Contrast (minimum). World Wide Web Consortium (W3C), 2021. https://www.w3.org/WAI/WCAG22/Understanding/contrast-minimum.html.

[2] Idris Adjerid, Alessandro Acquisti, Laura Brandimarte, and George F. Loewenstein. Sleights of privacy: framing, disclosures, and the limits of transparency. In *Proc. of the Workshop on Usable Security*, 2013.

[3] Fatemeh Alizadeh, Timo Jakobi, Alexander Boden, Gunnar Stevens, and Jens Boldt. GDPR Reality Check – Claiming and Investigating Personally Identifiable Data from Companies. In *Proc. of EuroS&PW*, 2020.

[4] Ryan Amos, Gunes Acar, Elena Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proc. of WWW*, 2021.

[5] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. Policylint: Investigating internal privacy policy contradictions on google play. In *Proc. of USENIX Security Symposium*, 2019.

[6] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with policheck. In *Proc. of USENIX Security Symposium*, 2020.

[7] Sajjad Arshad, Amin Kharraz, and William Robertson. Include me out: In-browser detection of malicious third-party content inclusions. In *Proc. of Intl. Conf. on Financial Cryptography*, 2016.

[8] Zineb Ait Bahajji and Gary Illyes. HTTPS as a ranking signal. Google Search Central Blog, 2014. https://developers.google.com/search/blog/2014/08/https-as-ranking-signal.

[9] Rebecca Balebako, Florian Schaub, Idris Adjerid, Alessandro Acquisti, and Lorrie Cranor. The impact of timing on the salience of smartphone app privacy notices. In *Proc. of the ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, 2015.

[10] Muhammad Ahmad Bashir, Sajjad Arshad, Engin Kirda, William Robertson, and Christo Wilson. How Tracking Companies Circumvented Ad Blockers Using WebSockets. In *Proc. of IMC*, 2018.

[11] Muhammad Ahmad Bashir, Sajjad Arshad, Engin Kirda, William Robertson, and Christo Wilson. A Longitudinal Analysis of the ads.txt Standard. In *Proc. of IMC*, 2019.

[12] Muhammad Ahmad Bashir, Sajjad Arshad, Wil Robertson, Engin Kirda, and Christo Wilson. A Longitudinal Analysis of the ads.txt Standard. In *Proc. of IMC*, 2019.

[13] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. Tracing information flows between ad exchanges using retargeted ads. In *Proc. of USENIX Security Symposium*, 2016.

[14] Muhammad Ahmad Bashir, Sajjad Arshad, and Christo Wilson. Recommended For You: A First Look at Content Recommendation Networks. In *Proc. of IMC*, 2016.

[15] Muhammad Ahmad Bashir and Christo Wilson. Diffusion of User Tracking Data in the Online Advertising Ecosystem. In *Proc. of PETS*, 2018.

[16] Harry Brignull. Dark patterns, 2019. https://www.darkpatterns.org/.

[17] Luca Bufalieri, Massimo La Morgia, Alessandro Mei, and Julinda Stefa. GDPR: When the Right to Access Personal Data Becomes a Threat. In *Proc. of ICWS*, 2020.

[18] Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110, 2021.

[19] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proc. of PETS*, 2016(4):237–254, 2016.

[20] Aaron Cahn, Scott Alfeld, Paul Barford, and S. Muthukrishnan. An empirical study of web cookies. In *Proc. of WWW*, 2016.

[21] Original Proposed CCPA Regulations, October 2019. https://www.oag.ca.gov/sites/all/files/agweb/pdfs/privacy/ccpa-proposed-regs.pdf.

[22] Chapter 20. California Consumer Privacy Act Regulations, 2020. https://govt.westlaw.com/calregs/Browse/Home/California/CaliforniaCodeofRegulations?guid=IEB210D8CA2114665A08AF8443F0245AD&originationContext=documenttoc&transitionType=Default&contextData=(sc.Default).

[23] First Set of Proposed Modifications to CCPA Regulations, February 2020. https://www.oag.ca.gov/sites/all/files/agweb/pdfs/privacy/ccpa-text-of-mod-redline-020720.pdf.

[24] Fourth Set of Proposed Modifications to CCPA Regulations, December 2020. https://www.oag.ca.gov/sites/all/files/agweb/pdfs/privacy/ccpa-prop-mods-text-of-regs-4th.pdf.

[25] Second Set of Proposed Modifications to CCPA Regulations, March 2020. https://www.oag.ca.gov/sites/all/files/agweb/pdfs/privacy/ccpa-text-of-second-set-mod-031120.pdf.

[26] Third Set of Proposed Modifications to CCPA Regulations, October 2020. https://www.oag.ca.gov/sites/all/files/agweb/pdfs/privacy/ccpa-text-of-third-set-mod-101220.pdf?

[27] AB-375 California Consumer Privacy Act of 2018, 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

[28] SB-1121 California Consumer Privacy Act of 2018, 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1121.

[29] Annotated Text of the California Privacy Rights Act, 2021. https://www.caprivacy.org/annotated-cpra-text-with-ccpa-changes/.

[30] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. Measuring Cookies and Web Privacy in a Post-GDPR World. In *Proc. of PAM*, 2019.

[31] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your Privacy... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proc of NDSS*, 2019.

[32] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proc. of CCS*, 2016.

[33] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. Large-scale readability analysis of privacy policies. In *Proc. of the International Conference on Web Intelligence*, 2017.

[34] Lesley Fair. Full disclosure. US FTC Business Blog, 2014. https://www.ftc.gov/news-events/blogs/business-

blog/2014/09/full-disclosure.

[35] Imane fouad, Nataliia Bielova, Arnaud Legout, and Natasa Sarafijanovic-Djukic. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. *Proceedings on Privacy Enhancing Technologies*, 2020(2):499–518, 2020.

[36] Stacia Garlach and Daniel D. Suther. 'I'm supposed to see that?' AdChoices Usability in the Mobile Environment. In *Proc. of HICSS*, 2018.

[37] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. How short is too short? implications of length and framing on the effectiveness of privacy notices. In *Proc. of the Workshop on Usable Security*, 2016.

[38] Helping advertisers comply with CCPA in Google Ads. Google Ads Help. https://support.google.com/google-ads/answer/9614122?hl=en.

[39] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. The Dark (Patterns) Side of UX Design. In *Proc. of CHI*, 2018.

[40] Colin M. Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. In *Proc. of CHI*, 2021.

[41] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. "it's a scavenger hunt": Usability of websites' opt-out and data deletion choices. In *Proc. of CHI*, 2020.

[42] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. An empirical analysis of data deletion and opt-out choices on 150 websites. In *Proc. of the Workshop on Usable Security*, 2019.

[43] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *Proc. of USENIX Security Symposium*, 2018.

[44] James Hercher. Google Strengthens Ads.txt Enforcement. ad exchanger, July 2018. https://adexchanger.com/ad-exchange-news/google-strengthens-ads-txt-enforcement/.

[45] Kashmir Hill. How an internet mapping glitch turned a random Kansas farm into a digital hell. Splinter News, 2016. https://splinternews.com/how-an-internet-mapping-glitch-turned-a-random-kansas-f-1793856052.

[46] Kashmir Hill. Why lost phones keep pointing at this Atlanta couple's home. Splinter News, 2016. https://splinternews.com/why-lost-phones-keep-pointing-at-this-atlanta-couples-h-1793854491.

[47] IAB CCPA Compliance Framework For Publishers & Technology Companies. Github, 2020. https://iabtechlab.com/standards/ccpa/.

[48] Carlos Jensen and Colin Potts. Privacy policies as decision-making tools: An evaluation of online privacy notices. In *Proc. of CHI*, 2004.

[49] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Faith Cranor, Shomir Wilson, Florian Schaub, and Norman Sadeh. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proc. of WWW*, 2020.

[50] Tobias Lauinger, Abdelberi Chaabane, Sajjad Arshad, William Robertson, Christo Wilson, and Engin Kirda. Thou shalt not depend on me: Analysing the use of outdated javascript libraries on the web. In *Proc of NDSS*, 2017.

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

[52] Timothy Libert. An automated approach to auditing disclosure of third-party data collection in website privacy policies. In *Proc. of WWW*, 2018.

[53] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The Privacy Policy Landscape After the GDPR. *Proceedings on Privacy Enhancing Technologies*, 2020(1):47–64, January 2020.

[54] Takaki Makino and Doantam Phan. Rolling out the mobile-friendly update. Google Search Central Blog, 2015. https://developers.google.com/search/blog/2015/04/rolling-out-mobile-friendly-update.

[55] Mariano Di Martino, Pieter Robyns, Winnie Weyts, Peter Quax, Wim Lamotte, and Ken Andries. Personal Information Leakage by Abusing the GDPR "Right of Access". In *Proc. of the Workshop on Usable Security*, 2019.

[56] Celestin Matte, Nataliia Bielova, and Cristiana Santos. Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *Proc. of IEEE Symposium on Security and Privacy*, 2020.

[57] Anthony D. Miyazaki. Online privacy and the disclosure of cookie use: Effects on consumer trust and anticipated patronage. *Journal of Public Policy & Marketing*, 27(1):19–33, 2008.

[58] Monica Nickelsburg. Why washington state could finally pass data privacy laws with a bill backed by the tech industry, January 2021. https://www.geekwire.com/2021/washington-state-finally-pass-data-privacy-laws-bill-backed-tech-industry/.

[59] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark Patterns after the GDPR: Scraping Consent Pop-Ups and Demonstrating Their Influence. In *Proc. of CHI*, 2020.

[60] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proc of NDSS*, 2019.

[61] Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James Graves, Fei Liu, Aleecia McDonald, Thomas Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh, and Florian Schaub. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30, August 2014.

[62] Sarah Rippy. Us state comprehensive privacy law comparison, February 2021. https://iapp.org/resources/article/state-comparison-table/.

[63] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. Can I Opt Out Yet?: GDPR and the Global Illusion of Cookie Control. In *Proc. of AsiaCCS*, 2019.

[64] Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, and Norman Sadeh. Automatic extraction of opt-out choices from privacy policies. In *Proc. of the AAAI Fall Symposium on Privacy and Language Technologies*, 2016.

[65] Jannick Sørensen and Sokol Kosta. Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In *Proc. of WWW*, 2019.

[66] Mukund Srinath, Shomir Wilson, and C. Lee Giles. Privacy at scale: Introducing the privaseer corpus of web privacy policies, 2020. https://arxiv.org/abs/2004.11131.

[67] David Stauss and Shelby Dolen. Virginia house passes consumer data protection act, February 2021. https://www.bytebacklaw.com/2021/02/virginia-house-passes-consumer-data-protection-act/.

[68] Peter Story, Sebastian Zimmeck, and Norman Sadeh. Which apps have privacy policies? an analysis of over one million google play store apps. In *Proc. of Annual Privacy Forum*, 2018.

[69] Jenny Tang, Hannah Shoemaker, Ada Lerner, , and Eleanor Birrell. Defining privacy: How users interpret technical terms in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021(3):70–94, 2021.

[70] Terms of service didn't read, 2021. https://tosdr.org/.

[71] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 4 Years of EU Cookie Law: Results and Lessons Learned. *Proceedings on Privacy Enhancing Technologies*, 2019(2):126–145, June 2019.

[72] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. A Study on Subject Data Access in Online Advertising After the GDPR. In *Proc. of Data Privacy Management, Cryptocurrencies and Blockchain Technology*, 2019.

[73] Privacy Online: A Report to Congress, 1998. https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-report-congress/priv-23a.pdf.

[74] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proc. of CCS*, 2019.

[75] Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, and Noah A. Smith. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Trans. Web*, 13(1), December 2018.

[76] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proc. of WWW*, 2016.

[77] Sebastian Zimmeck and Steven M. Bellovin. Privee: An architecture for automatically analyzing web privacy policies. In *Proc. of USENIX Security Symposium*, 2014.

[78] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. *Proceedings on Privacy Enhancing Technologies*, 2019(3):66–86, July 2019.

[79] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. Automated analysis of privacy requirements for mobile apps. In *Proc of NDSS*, 2017.

A Appendix

A.1 Crawling Methods

Our study relies on data collected from several web crawls. In this section, we describe the technical details of these crawls.

Our corpus of crawling targets comes from two sources: the Tranco list [60] and a list of known third-party A&A companies drawn from prior work [15]. This list contains effective second-level domains that must be translated to URLs for crawling. For this we used the following approach: our crawlers attempted to visit `https://domain/`, `https://www.domain/`, `http://domain/`, and `http://www.domain/` in that order, and accepted the first successful connection. As we note in § 2.3, we assume that the URL path / will (eventually) lead to a website's homepage. All of our crawlers follow redirects.

To collect static snapshots of websites' homepage HTML, we use the Python `requests` module. We configured `requests` to present a HTTP User-Agent string copied from Chrome.

To extract data from fully-rendered homepages (i.e., including all sub-resources and DOM modifications via JavaScript) we use the Python `requests-html`, `pyppeteer`, and `trio-chrome-devtools-protocol` modules to control Chrome via the Chrome Developer Protocol API. These modules all offer APIs that allow Python scripts to direct Chrome to load URLs into tabs, hook DOM events that occur within tabs, and inspect the DOM within tabs. After directing Chrome to load a given URL, our scripts wait 15 seconds for the webpage to completely load before inspecting the DOM. We use this timed-wait approach rather than relying on DOM events (e.g., `DOMContentLoaded` or `load`) since the latter are notoriously unreliable—these DOM events may fire before crucial page elements are loaded from third-parties, for example.

We used two methods to identify hyperlinks of interest for our study. We identified links in static HTML snapshots using regular expressions. Identifying links in static HTML snapshots is useful because this method is able to detect links that may be hidden or dynamically removed at runtime by JavaScript, e.g., due to client-side geofencing of DNSMPI links. We identified the DOM nodes containing links of interest in live webpages loaded within Chrome using CSS selectors and XPath queries. In both cases we paired our selectors with exclusion phrases (see § A.2) to eliminate false positives.

We extracted the display properties of DOM nodes containing DNSMPI, ToS, and PP links using the `.getBoxModel()` and `.getComputedStyleForNode()` methods offered by the Chrome Chrome Developer Protocol API. These methods take the viewport of the browser window and all available styling information into account when reporting the position, size, font styling, and color properties of a DOM node. We configured Chrome to use a window size of 1920×1080, i.e., the resolution of a standard 1080p monitor.

We used two methods to control the geolocation exposed by our crawlers to remote servers and JavaScript. First, we used the `.setGeolocationOverride()` method offered by the Chrome Chrome Developer Protocol API to set a latitude and longitude in California within Chrome. Thus, any JavaScript code that accesses the DOM `navigator.geolocation` property will believe the browser is in California. Second, we ran instances of our crawler on servers in Boston and in California; any website that relies on IP address geolocation will thereby infer the corresponding locations.

To collect inclusion trees from webpages, we used the DeepCrawling[23] tool. DeepCrawling is a NodeJS-based crawling tool that also leverages Chrome and the Chrome Developer Protocol API, and is specifically designed to record detailed information about the provenance of all sub-resources that are included within a webpage. DeepCrawl has been successfully used by several prior studies [7, 50], including those focused on online tracking and advertising [10, 11, 13–15], which is also why we adopt it for this study.

Unfortunately, the inclusion trees produced by DeepCrawling only allow us to causally identify the source of resource inclusions—they do not allow us to identify the source of arbitrary DOM modifications (unless those modifications cause a resource to be loaded over the network). Thus, we cannot use inclusion trees to identify which scripts are responsible for client-side hiding of DNSMPI links (see § 4.3).

A.2 Hyperlink Detection Phrases

Our crawler used the following terms and phrases to identify potential ToS and PP hyperlinks, subject to the following list of exclusion phrases.

ToS Phrases. "terms", "user agreement", "service agreement", "conditions of use", "terms of usage".

PP Phrases. "privacy notice", "privacy policy", "privacy & cookies".

Exclusion Phrases. "preferences", "terms of sale", "login", "terms and conditions apply", "accessibility", "your data in search", "shield", "promo", "campaign", "deal", "ad choice", "january", "february", "march", "april", "may", "june", "july", "august", "september", "october", "november", "december", "archive", "previous", "versions", "settings".

A.3 CCPA/CPRA Compliance by Third-Parties

Determining whether the relationship between a first- and third-party constitutes sharing and/or selling data under the CCPA and CPRA is challenging, even when the third-party in question is an A&A company. This challenges arises from the "service provider" exemption in the CCPA and CPRA: CA Civ Code §1798.140(v) defines "service providers" as businesses that may hold and process PI from a first-party, but are forbidden from selling or sharing that data with other third-parties, or using the data for their own commercial purposes. An uncontroversial example of a service provider is a web hosting company: by vestige of their business model, they will necessarily observe PI from people visiting their customers' websites (e.g., IP addresses), but their business does not (typically) involve sharing, selling, or otherwise monetizing this data.

In contrast, Google's tracking and advertising products present a more complex example of a service provider. By default, these products collect and monetize PI from first-party publishers, but publishers can change their settings to put these products in "restricted data processing" mode [38]. When this mode is enabled, Google's products collect and store less data about people, and many forms of targeted advertising are disabled (but other forms of advertising, e.g., contextual, remain). Google argues that when "restricted data processing" mode is enabled their products become CCPA compliant, even if the publisher in question has not offered their users a chance to opt-out of the data col-

[23] https://github.com/sajjadium/DeepCrawling

lection, because they are now operating as a service provider for the publisher.

The IAB Tech Lab has produced a specification and API (the *US Privacy String*) [47], that provides a generic mechanism for first-party publishers to label data flows to third-parties as falling under the service provider exemption. This specification is similar to the Transparency & Consent Framework (TCF) developed by IAB Europe for facilitating GDPR compliance. Google supports the US Privacy String API, in addition to other A&A companies.

In summary: a first-party website may not need to include a DNSMPI link on their homepage if they (1) state in their privacy policy that they do not sell data and (2) use APIs like those discussed above to communicate to their third-party partners that they should act as service providers, rather than data processors. In this study, we do not attempt to identify third-party HTTP requests that contain meta-data signifying that service provider mode is in effect; detecting and interpreting these signals in non-trivial [56] and thus we leave this as future work.