Privacy Lost and Found: An Investigation at Scale of Web Privacy Policy Availability

Mukund Srinath
mukund@psu.edu
Pennsylvania State University
University Park, PA, USA

Soundarya Sundareswara
soundarya.sundareswara@gmail.com
Pennsylvania State University
University Park, PA, USA

Pranav Venkit
pranav.venkit@psu.edu
Pennsylvania State University
University Park, PA, USA

C. Lee Giles
clg20@psu.edu
Pennsylvania State University
University Park, PA, USA

Shomir Wilson
shomir@psu.edu
Pennsylvania State University
University Park, PA, USA

ABSTRACT

Legal jurisdictions around the world require organisations to post privacy policies on their websites. However, in spite of laws such as GDPR and CCPA reinforcing this requirement, organisations sometimes do not comply, and a variety of semi-compliant failure modes exist. To investigate the landscape of web privacy policies, we crawl the privacy policies from 7 million organisation websites with the goal of identifying when policies are unavailable. We conduct a large-scale investigation of the availability of privacy policies and identify potential reasons for unavailability such as dead links, documents with empty content, documents that consist solely of placeholder text, and documents unavailable in the specific languages offered by their respective websites. We estimate the frequencies of these failure modes and the overall unavailability of privacy policies on the web and find that privacy policies URLs are only available in 34% of websites. Further, 1.37% of these URLs are broken links and 1.23% of the valid links lead to pages without a policy. Further, to enable investigation of privacy policies at scale, we use the *capture-recapture* technique to estimate the total number of English language privacy policies on the web and the distribution of these documents across top level domains and sectors of commerce. We estimate the lower bound on the number of English language privacy policies to be around 3 million. Finally, we release the CoLIPPs Corpus containing around 600k policies and their metadata consisting of policy URL, length, readability, sector of commerce, and policy crawl date.

CCS CONCEPTS

• **Information systems** → **Web mining**; • **Security and privacy** → **Privacy protections**.

KEYWORDS

privacy, privacy policy, capture-recapture, policy availability

DocEng '23, August 22–25, 2023, Limerick, Ireland

ACM ISBN 979-8-4007-0027-9/23/08…$15.00
https://doi.org/10.1145/3573128.3604902

ACM Reference Format:
Mukund Srinath, Soundarya Sundareswara, Pranav Venkit, C. Lee Giles, and Shomir Wilson. 2023. Privacy Lost and Found: An Investigation at Scale of Web Privacy Policy Availability. In *ACM Symposium on Document Engineering 2023 (DocEng '23), August 22–25, 2023, Limerick, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3573128.3604902

1 INTRODUCTION

Privacy policies are legal documents that organizations use to disclose how they collect, analyze, share, and secure their users' personal data. This disclosure is made mandatory in a number of jurisdictions through laws such GDPR (General Data Protection Regulation) in the European Union and a number of state laws in the United States[1], such as CCPA (California Consumer Privacy Act) amended to CPRA (California Privacy Rights Act) and VCDPA (Virginia Consumer Data Protection Act). More than 182 legal jurisdictions around the world have some form of privacy laws, and globally, the number of privacy laws has been growing exponentially since the 21st century [12]. These laws work under the principle of *notice and choice*. *Notice* is a presentation of terms (i.e., the privacy policy) and *choice* is an action signifying the acceptance of those terms (i.e., continuing to use the site). Therefore, users often accept the privacy policy of an organization by simply using their offered services. It is thus assumed that users have understood and accepted an organization's privacy practices by reading its privacy policy before using the services offered.

Across the world, privacy policies tend to be the primary and often the only source of information regarding what happens to users' personal information online. The GDPR (Articles 12-14), explicitly requires transparent disclosure of information, accurately reflecting data practices. The CPRA further formalizes this requirement, making it mandatory for organizations that exceed a $25M annual revenue or process information of more than 50k users to post a privacy policy and update it at least once a year [3]. However, policy regulators are often overwhelmed by the number of privacy policies on the web. Policy regulators even in privacy-friendly jurisdictions, such as the EU, rely on user complaints to investigate privacy practices [36] while others (such as the United States) rely on organizations to self-certify their compliance[2] and only investigate when a privacy policy is at odds with real-world privacy

[1] https://iapp.org/resources/article/us-state-privacy-legislation-tracker/
[2] https://www.privacyshield.gov/Program-Overview

practices. It is therefore essential that all organizations make their privacy policies easily accessible to users so that they can make an informed decision about their privacy online.

However, studies have found that privacy policies are often time-consuming to read and difficult to understand [11, 24]. Studies suggest that a reader would need to be educated at a college level to be able to understand an average privacy policy [10, 24]. Further, an individual would have to spend an average of about 154 hours per year to skim through all the privacy policies from websites they visit [23]. This issue is compounded by the fact that policies are often unavailable. To this end, we investigate the lack of availability of policies by studying the frequency of various failure modes in privacy policy availability, due to issues such as broken links, empty content, placeholder text and language issues. Our analysis of trends in failure modes of privacy policies will throw light on the issues in online privacy landscape thereby motivating further study. Lack of availability of privacy policies is especially worrying since it signifies a breakdown of *notice and choice* and therefore warrants careful investigation. Next, we undertake the task of estimating the lower bound on the total number of privacy policies on the web using the capture-recapture (Lincoln-Petersen estimator) technique [4] in order to study the trends in distribution of policies on the internet. We estimate the distribution of website privacy policies in terms of sectors of commerce and the most popular top level domains (TLDs).

To undertake this study, we begin with a collection of company websites and their corresponding sectors of commerce from LinkedIn. We then crawl the websites' landing pages and their corresponding privacy policies (if one exists) and take a note of the various error modes we encounter. Additionally, we manually sample and investigate 500 candidate privacy policy pages. We find users are likely to find a privacy policy in less than 34% of websites due a number of failure modes such as absence of policy, dead links, empty content, placeholder text and language discrepancies. This brings into question the efficacy of the notice and choice as a framework for user privacy online. Additionally, we release the corpus of 600k privacy policies along with their metadata consisting of policy URL, length, readability calculated using the Flesch Kincaid grade level formula [16], sector of commerce, and policy crawl date. Next, we estimate the total number of privacy policies on the web and their distribution based on sectors of commerce and various top level domains thereby showing light on the trends of policies online. We find the lower bound on the number of English language privacy policies on the web is around 3 million based on the capture-recapture estimation technique.

As such, we make the following contributions:

- Create and share a corpus of 600k privacy policies with rich metadata information[3].
- Provide a detailed analysis on various anomalies found related to availability of privacy policies on the web.
- Estimate a lower bound for the total number of English language privacy policies on the web using the capture-recapture technique.
- Discuss trends in the availability of privacy policies based on sector of commerce and TLD information.

[3]https://privaseer.ist.psu.edu/data

2 RELATED WORK

In this section we first list the various privacy policy corpora available and the natural language processing techniques applied on them in order to make policies more comprehensible. Next, we discuss literature studying the issues of use and availability of privacy policies online. Finally we discuss literature related to the estimation of the number of privacy polices on the web by looking into the applications of the capture-recapture technique.

2.1 Privacy Policy Corpora

Natural language processing (NLP) techniques have been applied on prior corpora of policies in order to ease users' burden of reading them. Small corpora of policies (up to a few thousand) have been examined using question-answering systems [27], chatbots [13], and information extraction techniques [28, 29, 32] with promising results. Nevertheless, the variation in language use and coverage of policies across different jurisdictions and sectors of commerce indicates that these small corpora may not be sufficiently robust for real-world use cases. More recently, larger corpora, such as the PrivaSeer Corpus [30, 31, 33], and the Princeton-Leuven Longitudinal Corpus of Privacy Policies [2], both containing over one million privacy policies, have enabled more advanced language technologies to comprehend privacy policies. However they do not share important metadata information such as readability score, length, sector of commerce information, and last updated date. We release the CoLIPPS Corpus containing about 600k policies along with its metadata.

2.2 Privacy Policy Issues

Previous investigations on the insufficiency of privacy policies have mainly concentrated on their complex language and the reasons behind users' reluctance to read them. Nonetheless, scrutiny of privacy policies has revealed a concerning silence on vital consumer-related practices, such as the collection and utilization of sensitive information, as well as the integration of tracking data with personally identifiable information. Many of these policies have been found to lack full compliance with self-regulatory guidelines [8, 14]. Moreover, analysis of policies from 16 of the globe's largest internet and telecommunications companies has indicated that vague or unclear language impedes users' comprehension of company practices, hampering informed decision-making on products and services [18]. The implementation of the General Data Protection Regulation (GDPR) in May 2018 has played a pivotal role in ensuring online privacy rights for users [17]. Studies have further investigated the factors that correlate with GDPR compliance practices within organizations by examining the corresponding privacy policies [1].

In this work, we investigate a more fundamental question about trends in the ability of access to these documents. Another similar line of research involved identifying mismatches between user expectations and privacy practices stated in privacy policy documents [26]. This work described the potential of highlighting unexpected practices in websites thereby helping users to make better privacy decisions. Similarly, Sunyaev et al. assessed the availability, scope, and transparency of privacy policies of mobile health apps on iOS

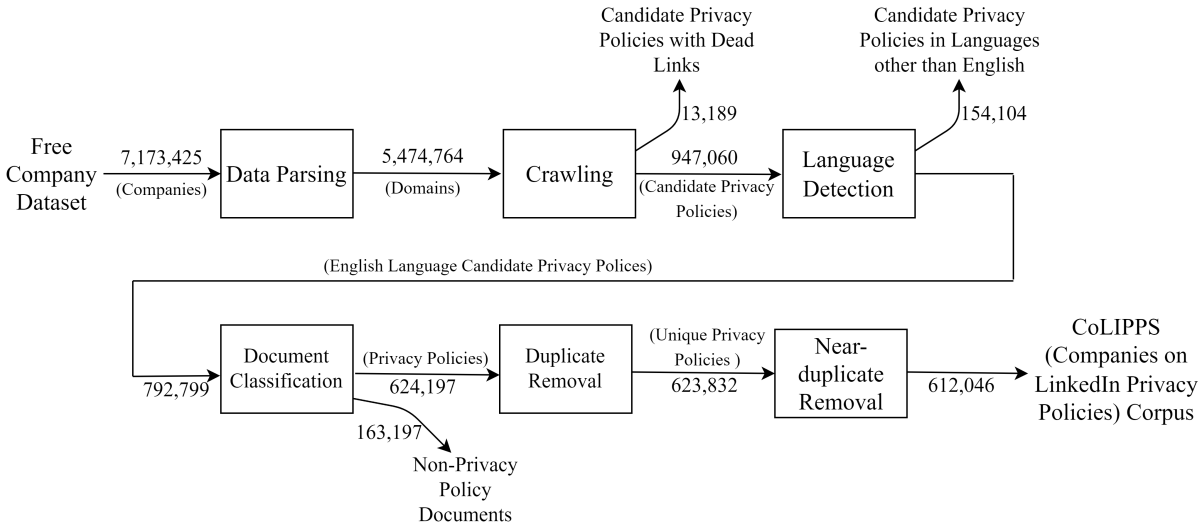

Figure 1: Processing pipeline for document collection and classification

and Android and showed that privacy policies have poor transparency and availability rates. They found that of the 600 most commonly used apps, only 183 (30.5%) had privacy policies [35].

2.3 Document Estimation

The capture-recapture technique, also known as the Lincoln-Petersen estimator or mark-recapture, is a solution to the problem of estimating the size of a population when it is not feasible to count every individual item within it. This technique has traditionally been utilized in ecology to estimate animal populations [4] and in epidemiology to estimate disease prevalence [6]. For example, the capture-recapture technique was employed by Bohning et al. to estimate the number of COVID-19 infections [5]. In the realm of information security, Weaver et al. utilized this technique to gauge the extent of phishing activity on the internet [38]. In information retrieval, researchers have applied this technique to estimate the size of the world wide web [20] and the coverage of major search engines [9, 19]. Khabsa and Giles [15] estimated the number of scholarly documents available on the web and the size of Google Scholar using this method. In this paper, we estimate the number of English language privacy policies on the web in order to inform trends in the online privacy landscape.

3 PRIVACY POLICY CORPUS CREATION

In order to create the privacy policy corpus, we begin with a list of company domains from the Free Company Dataset[4] provided by People Data Labs. The Free Company Dataset is a collection of over 7 million global companies containing fields such as name, domain, year founded, industry, size range, locality, country, LinkedIn URL, current employee estimate, and total employee estimate. Figure 1 shows the processing pipeline for extracting privacy policies from this dataset. This pipeline is largely based on the work of Srinath et al. [31], Sundareswara et al. [34] which was designed to harvest privacy policies from the web.

We extracted a list of company domain URLs from the dataset and obtained about 5.5 million unique URLs. We then used Scrapy[5] to crawl the URLs. About 1.5 million domain URLs returned an error while 4 million were successfully crawled. Privacy laws around the world have created an industry standard to include a link to the privacy policy in the footer of the website landing page. Additionally, we found that privacy policy URLs tend to have the word 'privacy' or the words 'data' and 'protection' in them. Thus we extracted candidate privacy policy URLs from website domain pages by searching for HTML *href* attributes containing the words *privacy* or the words *data* and *protection*. We refer to this as our **URL selection criteria**. The majority of the successful domain URLs did not produce a successful potential privacy policy URL since they either did not have a privacy policy, or none of the the hyperlinks in the page satisfied our URL selection criteria. We crawled a total of about 950k web pages that satisfied our URL selection criteria. We refer to these web pages as **candidate** privacy policy documents. Next, we put the candidates policies through LangID, a language identification tool [21]. We used LangID due to its high accuracy over a range of languages and domains. LangID accepts a segment of text and returns the identified language and is capable of detecting 97 languages. We retained the 790k candidates were identified to in English for further analysis and discarded the rest.

In order to separate privacy policies from other types of documents that only satisfied our URL selection criteria, we used a machine learning approach. We uniformly randomly sampled 1,600 English documents and labelled them to be either a privacy policy or not. We then trained a random forest model with features extracted from the policy document and URL. We separately tokenized the words in the policy document and policy URL and removed stop words. We then calculated the term frequency-inverse document

[4]https://docs.peopledatalabs.com/docs/free-company-dataset

[5]https://scrapy.org/

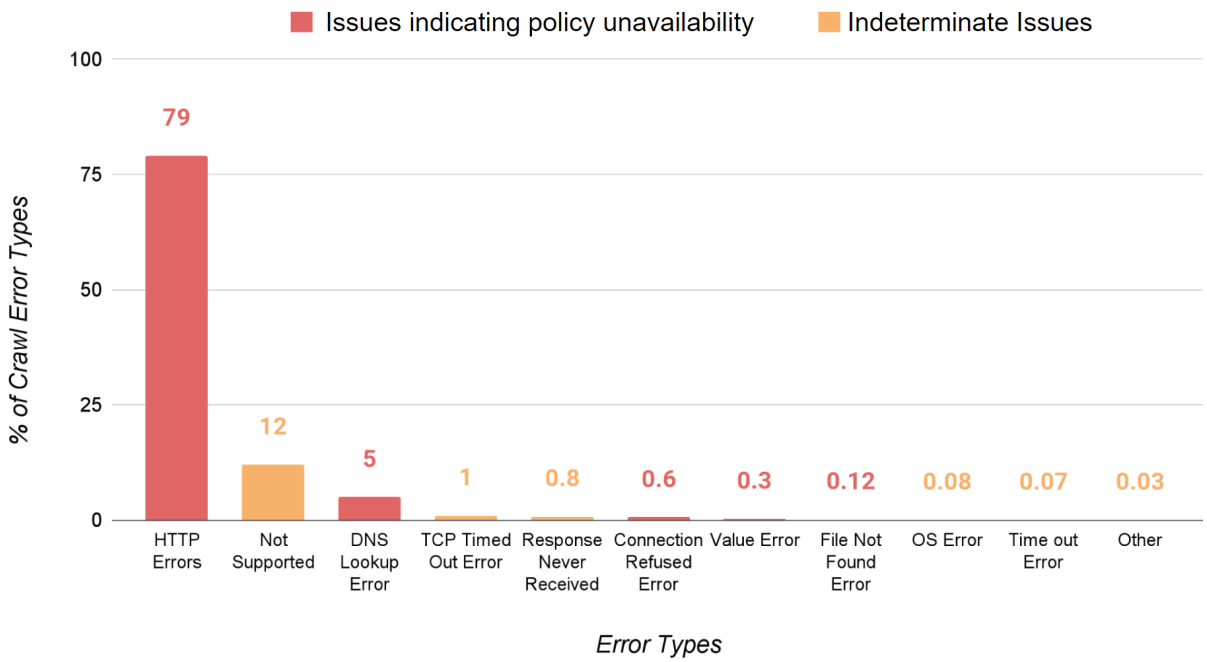

Figure 2: Types and Frequencies of crawling errors

frequency (tf-idf) for each policy. We divided the labelled documents into train, validation and test sets in the ratio 3:1:1. After tuning the hyperparameters using the validation set and a 5-fold cross-validation, the precision, recall and F1 scores were found to be 0.96, 0.97 and 0.96 respectively. We ran the trained random forest classifier on the English language candidate privacy policies and found that 625k documents were classified as privacy policies while the other 165k were not privacy policies.

We removed duplicates and near-duplicates from the 625k documents classified as privacy policies. Duplicates were removed by hashing the downloaded privacy policy files and discarding privacy policies that generated a duplicate hash. To remove the near-duplicates, we first grouped privacy policies belonging to the same domain. We then applied the simhash [7] algorithm on shingles of size three for each privacy policy. Hamming distance between each pair of privacy policies within the same domain was calculated using the simhashes and near-duplicates were filtered out based on a Hamming distance threshold [22]. We obtained 612,046 unique English language privacy policies from the Free Company Dataset. We denote this set of privacy policies as the CoLIPPs (Companies on LinkedIn Privacy Policies) Corpus.

3.1 Evaluation of URL Selection Criteria

In order to estimate the accuracy of our URL section criteria, we manually examined a random sample of 500 company URLs, from the 4 million successfully crawled domain URLs. Table 1 lists our observations from this sample. We found that 330 domains did not contain a hyperlink to a privacy policy, while 170 domains listed one on their landing page. Among the 170 policy URLs found, 106 contained the chosen keywords (*privacy* or *data* and *protection*) as is, while 42 had the chosen keywords (as well as the privacy policy itself) in other languages. Additionally, 22 websites displayed their policies with JavaScript in HTML elements such as a dialog box. Out of these, 17 documents were in English.

Our work focuses on collecting English language privacy policies. Therefore, while estimating the accuracy of our URL section criteria, we exclude the websites that were in languages other than English. Thus, given that 106 English language websites had the chosen keywords in their privacy policy URL, out of a total of 123 English language websites, we estimate that our URL selection criteria captured the privacy policies of 86.17 ± 7.38 % of English language websites. We also found that only 34 ± 4.15% of websites in the sample had a privacy policy hyperlink on the landing page irrespective of language.

Table 1: Observations on a random sample of 500 company URLs from successfully crawled domain URLs

Observation		Number
Websites without a privacy policy hyperlink		330
Websites with a privacy policy hyperlink	Chosen keywords in privacy policy URL	106
	Translated versions of chosen keywords in privacy policy URL	42
	Privacy policy displayed using other HTML elements	English: 17 Others: 5
	Total	**500**

4 POLICY AVAILABILITY ANALYSIS

In this section, we examine three distinct failure modes of privacy policy availability, namely dead/broken links, language discrepancies, and non-privacy policy documents.

4.1 Dead Links to Privacy Policies

Figure 2 shows error types that had at least 10 instances when crawling the URLs that satisfied our selection criteria. The figure shows the percentage of each error type that were issued by the server in response to a request made by the crawler. The error Types *HTTP Error*, *DNS (Domain Name Server) Lookup Error*, *Connection Refused Error*, *Value Error*, and *File Not Found Error* indicate unavailability of privacy policies, since these error types indicate issues with the availability of a web page. The error types *Not Supported* (crawler not being able to perform an action on HTML elements such as JavaScript pop-ups), *TCP Timed Out*, *Response Never Received*, *OS Error* and *Timeout Error* could be due to any number of issues with crawling and are therefore not indicative of unavailable pages.

13,261 URLs failed with *HTTP Errors*, with a majority returning *404 (Page Not Found)* and *500 (Internal Server Error)*, both indicative of an error in configuring the requested page. 764 URLs failed with *DNS Lookup Errors*; since the URLs were obtained from successfully crawled domains, a DNS error suggests that the URLs were invalid. 95 instances of *Connection Refused Errors* were observed; manual inspection revealed that these errors were either due to URLs referencing *localhost* or due to incorrectly configured ports. 51 URLs had *Value Errors*; manual examination showed that they were comprised of invalid URLs. 17 links returned *File Not Found Errors*, which suggested that they had invalid configurations. Summing all the errors due to the above error types, we found that 14,188 URLs could possibly be due to unavailable web pages.

To investigate the nature of the 14,188 error URLs that were likely due to unavailable privacy policies, we randomly sampled 100 and found that 98 URLs had all the markers of a privacy policy URL i.e. containing the term 'privacy', present in the footer of the landing page. This suggests that a conservative estimate of 13,189 error URLs (95% confidence inter) did not redirect to a valid privacy policy page. We therefore estimate that 13,189 out of all the candidate policies or 1.37% of all privacy policy URLs irrespective of language are dead links. It should be noted that this is a conservative estimate, since we calculate the percentage of dead links with respect to the collection of candidate privacy policy URLs not actual privacy policy URLs.

4.2 Natural Language Discrepancies

With many websites being offered in multiple languages, we investigated whether the websites which were offered in multiple languages also offered their privacy policy in those languages. We identified inconsistencies where privacy policy text was unavailable in certain languages. For example, in one particular case the website was displayed in English with options to switch to Spanish and Portuguese. When we attempted to access the privacy policy in the English version, we were directed to the document in Spanish. This behaviour certainly does not help users not literate in Spanish, who would be unable to read the document without translation services.

To estimate such inconsistencies, we randomly sampled 250 domain URLs from the 4 million domains which were successfully crawled. 31 out of the 250 domains were available in multiple languages, with English always being one of the languages in which the website was offered. 13 domains presented inconsistent user experience by not displaying the privacy policy document in a preferred language set on the landing page. Only 10 domains offered the privacy policy in all the languages in which the rest of the website was offered, while the remaining 8 did not have a privacy policy in any language. Thus, we find that 41.9 ± 17.35% of the websites that are offered in multiple languages have privacy policy language inconsistencies, and in general we estimate that 5.2 ± 2.4% of all the websites (where at least one of the languages is English) display this type of language inconsistency with respect to privacy policies.

Additionally, we found that about 2,858 were in Latin, a language that is unsuitable for legal notice and choice in any modern jurisdiction. We informally examined the contents of these documents and found most consist of placeholder texts, such as *Lorem ipsum* [39].

4.3 Non-Privacy Policy Documents

20.71% (163,049) of candidate privacy policy documents in English were classified as non-privacy policy documents. Manual inspection of these documents revealed that they included documents with valid privacy policy URLs. To identify the reason for the classification, we examined the content of these documents. We found that a number of valid privacy policy URLs led to web pages that had empty privacy policies, i.e., they had the heading *Privacy Policy* but did not have any content.

To estimate the rate of occurrence of empty privacy policies, we randomly sampled 500 non-privacy policy documents and manually inspected their content. 6.8 ± 2% (34) of documents in the random sample had empty content on the privacy policy page suggesting that 7,753 (considering the lower bound) documents out of the 163,049 documents classified as not a privacy policy are in reality privacy policy web pages without any content. We estimate that on average, 1.23% of all valid English language privacy policy URLs have empty content.

5 PRIVACY POLICY ESTIMATION

To estimate the number of English language privacy policies on the web, we used the Lincoln-Petersen estimator [25], commonly used to estimate wildlife species population in ecological studies. In this technique, a sample of animals from a population is captured, marked, and released back into the wild. Then, a second set of animals from the same population is captured and the number of previously marked animals are counted. The total number of animals in the population is then estimated based on the intuition from Bayes rule that the ratio of number of animals captured in the first sample to the total population is equal to the ratio of marked animals recaptured to the number of animals captured in the second sample, given that the two samples are random and independent.

We used privacy policies collected from two separate crawls as our two samples. We used the PrivaSeer Corpus [31] as the first sample and the CoLIPPs Corpus that we created as as the second sample. The PrivaSeer Corpus is a collection of 1,005,380

Table 2: Top 10 TLDs with the greatest number of privacy policy documents.

TLD	Privaseer Corpus	CoLIPPs Corpus	Overlaps	Estimate	% Share
.com	632,971	387,377	130,168	1,883,707	63.1
.uk	113,515	78,353	27,149	327,608	10.99
.org	55,133	34,832	13,630	140,895	4.73
.au	38,420	30,856	10,329	114,839	3.85
.net	28,718	11,830	3853	88,588	2.97
.ca	16,123	9552	3227	47,724	1.6
.info	8049	576	173	26,799	0.89
.in	6251	3843	928	25,886	0.86
.ie	6884	4288	1509	19,505	0.65
.de	6321	2835	929	19,290	0.64

English language privacy policies crawled in July 2019 using URLs extracted from CommonCrawl. CommonCrawl releases monthly crawl archives which provide a representative sample of the web[6]. Since CommonCrawl monthly archives are collected independent of the Free Company Dataset, which we used to build the CoLIPPs Corpus, we assume that the PrivaSeer Corpus and the CoLIPPs Corpus represent independent samplings.

To estimate the number of English language privacy policies on the web, we extracted all the URLs from the PrivaSeer Corpus (n_1) and counted them as marked samples. We then extracted all the URLs from the CoLIPPs Corpus (n_2). Next, we counted the number of recaptured (marked in the first sample) URLs i.e. the URLs that are present both in the PrivaSeer Corpus and the CoLIPPs Corpus (m) and estimated the total number of English language privacy policies on the web (N), including a confidence interval, using the following formulae [25].

$$N = \frac{n_1.n_2}{m} \tag{1}$$

$$100(1-\alpha)\% = \frac{n_1.n_2}{m} \pm z_{1-\alpha}\sqrt{\frac{n_1.n_2.n_{12}.n_{21}}{m^3}} \tag{2}$$

where n_{12} is the number of objects captured in the first sample but not the second and n_{21} is the number of objects captured in the second sample but not the first.

The Lincoln-Petersen technique makes three assumptions for estimation:

(1) That the population is closed to additions and deletions: This means that between the two samples, no new objects can be added or removed. The Lincoln-Petersen technique, however allows for unmarked additions, and deletions if they occur randomly with respect to marked and unmarked objects.
(2) That marks are not lost or altered such that an object marked in the initial sample could not be re-identified if later recaptured.
(3) That all objects are equally likely to be captured.

With respect to the assumption (1), although the number of privacy policies on the web is not a closed population, all privacy policies added after the first sample were not previously crawled, therefore we can consider them to be unmarked. We can also assume that deletions occurred randomly between marked and unmarked privacy policies due to the constant flux in the number of websites and therefore privacy policies on the web. Thus, we can assume that assumption (1) is satisfied.

We tailored our marking technique to satisfy assumption (2). We marked the privacy policies based on the domain name found in its website URL, i.e., if the domain name of the privacy policy URL in the PrivaSeer corpus was found in the CoLIPPs Corpus, then it was counted as a recapture even if the URL paths were different. We followed this technique since the two corpora were crawled seven months apart, during which several privacy policy URL paths within a website might have changed. In order to use domain names as markers, we made the additional assumption that all domains have at most one privacy policy, since the vast majority of websites on the internet have only one valid current privacy policy.

With respect to assumption (3), both the selected corpora use a URL selection criteria to filter privacy policy candidates. The URL selection criteria excludes all URLs which do not have the word *privacy* or the words *data* and *protection* in them. We estimated that this selection criteria excludes a maximum of about 21.2% (from the evaluation of the URL selection criteria in Section 3) of the English language privacy policies on the web. Thus, compared to a random privacy policy online, it is more likely that the objects marked in the first sample are recaptured in the second. We also acknowledge that the sources for both the corpora might have their own biases. For example, CommonCrawl, the source for the PrivaSeer Corpus, create their monthly archives based on seed URLs from various sources and uncrawled domains from previous archives. Thus, the domains crawled might be biased towards the original seed URLs based on the method of randomization. Similarly, the CoLIPPs Corpus was derived from a corpus created by crawling LinkedIn. Thus the domains in the corpus might be biased towards organisation/company domains. Although this bias might partially be offset by the size of the corpora, it is unlikely to be completely removed. This would therefore cause an underestimation of the total number of privacy policies on the web. Since this bias cannot be completely avoided, we argue that our estimate is a lower bound on the number of privacy policies on the web.

Based on the PrivaSeer Corpus, we found that $n_1 = 1,005,380$. Based on the CoLIPPs Corpus, we found that $n_2 = 612,046$ and $m = 206,417$. Using equation (1), we estimated that the lower bound on the total number of privacy policies on the web is $2,981,047$.

[6]https://commoncrawl.github.io/cc-crawl-statistics/

Using equation (2), we estimated that the 95% confidence interval on the estimate is 9, 333.

5.1 Top Level Domain Analysis

In addition to estimating the lower bound of the number of privacy policies on the web, we estimated the number of privacy policies in each top level domain (TLD). Similar to our earlier estimate, we counted the number of common privacy policies between the PrivaSeer Corpus and the CoLIPPs Corpus based on the domain name and separated them by TLD. We then applied the Lincoln-Petersen technique to estimate the number of privacy policies in each TLD. Table 2 shows the top 10 TLDs with the largest number of estimated privacy policy documents. In this table, *Overlaps* and *Estimate* refer to the number of overlaps between the PrivaSeer Corpus and the CoLIPPs Corpus and the total estimated count of URLs in each TLD respectively. From this table, we can see that .com domains have the overwhelming majority of privacy policies, followed by .uk and .org domains. We also note that a number of country level TLDs, such as .uk, .au, .ca, .ie, .de, and .in contain a large number of privacy policies. A number of TLDs that appear in the top 10 registered TLDs on the internet [37] such as .tk, .cn, .icu, .nl and .ru do not appear in the top ten TLD with the greatest number of privacy policies. This could be due to the fact that we only capture English language privacy policies while a large percentage of websites in .cn. .nl and .ru TLDs are likely in Chinese, Dutch and Russian respectively. The differences in privacy policy frequencies in various country level TLDs could also be due to the privacy laws in the respective countries. Additionally, about 37% of the total number of registered domains on the internet have a .com TLD, however, 63% of the English language privacy policies are contained within .com TLD. This phenomenon might be explained by the fact that commercial establishments might prefer to host their websites on the .com TLD.

Figure 3 shows the TLDs with the highest number of English language policies. Each bar in the figure depicts the percentage of domains with and without a privacy policy for that TLD. Here we can see that country level TLDs with predominantly English speaking population have the highest ratio of domains with polcies to domains without policies, namely .uk (United Kingdom) and .au (Australia) and .ie (Ireland). We can see that .de, the German TLD contains contains a very low ratio of domains with policies to domains without policies, providing further evidence to our hypothesis. We see the the .in (India) TLD in spite of having a large population of English speakers and websites in English contain a much smaller ratio of domains with polices to without policies. This is likely due to the fact that unlike USA and the member states of the EU, India does not have a well defined law requiring privacy policies on websites.

Table 2 can also be used to identify bias while estimating population sizes. By breaking up the estimates into component categories [25] and analysing TLD estimates separately, we can identify instances of bias in the data. For example, we see TLD .info has a large difference between the number of overlaps and estimates. This might be due to the fact that one of the corpora disproportionately captured instances that TLD, thereby causing a positive

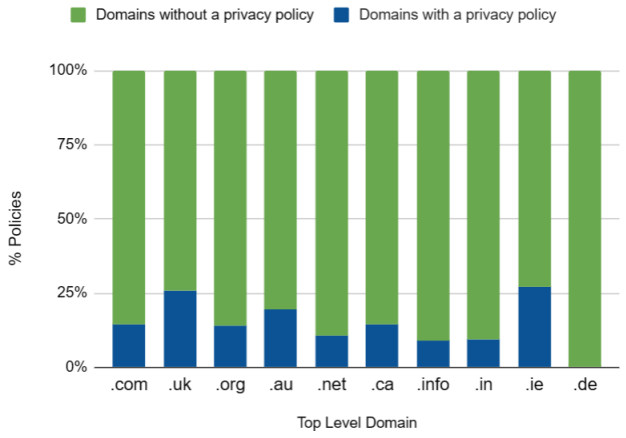

Figure 3: Comparison between the %domains containing a privacy policy and %domains not containing a privacy policies for the top ten TLDs with the most English langauge policies

estimation bias, i.e., overestimation of the number of privacy policies. The CoLIPPs Corpus was created from URLs in LinkedIn. It is reasonable to assume that not many companies would host their website on a .info domain. Thus, .info being underrepresented in the CoLIPPs Corpus might lead to an artificially low number of overlapping URLs between the corpora, causing an over-estimation of the number of privacy policies for the .info TLD. We can also see that TLDs which make up a large percentage of the total estimate such as .com, .uk and .org do not show any obvious signs of bias.

5.2 Sector of Commerce Analysis

The sector of commerce is an important factor to consider in the context of privacy policies. Not only do different sectors of commerce differ in trends in privacy practices and policy availability, they also collect different categories of personal information types from the user. For example, financial institutions might typically require a user's social security number, but this is not the case for e-commerce companies. The Free Company Dataset, from which we built the CoLIPPs Corpus, maps company websites to a set of 148 unique industries. The company categorization scheme employed in the Free Company Dataset comes from the categorization scheme followed by LinkedIn (companies self-report the industry to which they belong). We consolidated the 148 industries categories into 11 sectors of commerce. We make both the sector of commerce labels and industry labels available as part of the CoLIPPs corpus metadata.

Figure 4 shows the count privacy policies in each sectors of commerce in the CoLIPPs Corpus. We can see that the sectors *finance, marketing and human resources* and *information technology and electronics* having the greatest segments with 17% and 14% of the privacy policies respectively. Figure 5 shows the percentage of domains containing a privacy policies and the percentage of domains not containing a privacy policy for all the sectors of commerce. From the figure we can see that the medical sector has the highest ratio of domains containing policies to those that do not. On the

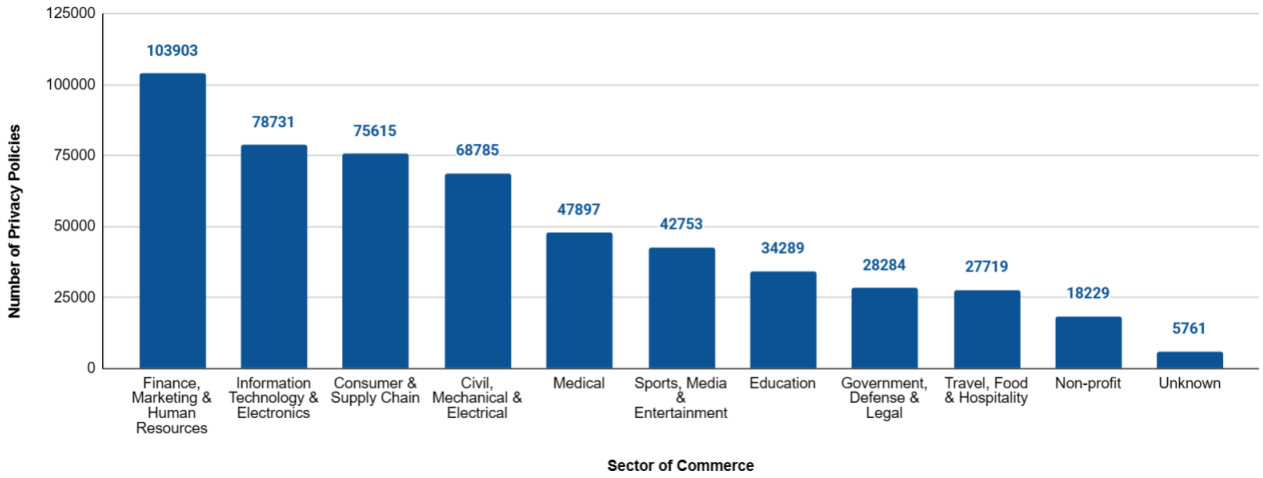

Figure 4: Distribution of the number of privacy policies in each sector of commerce in the CoLIPPs Corpus

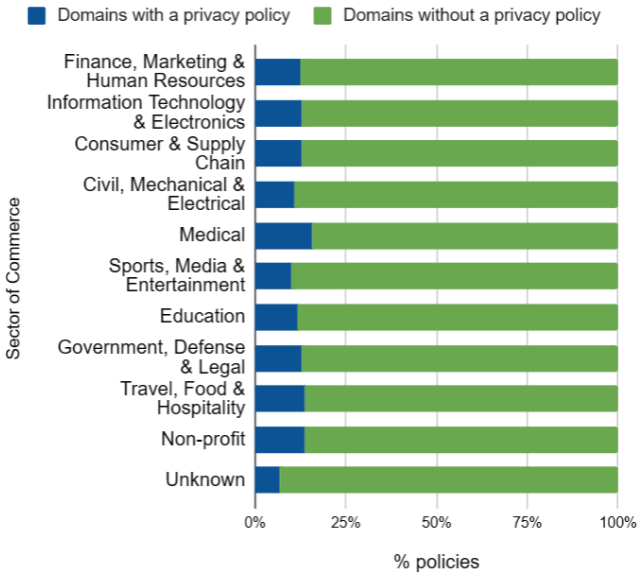

Figure 5: Comparison between the %domains containing a privacy policy and %domains not containing a privacy policies for all sectors of commerce

other hand, the sports media and entertainment sector has the lowest. We also see that there there is a general trend that consumer facing sectors that could potentially serve as a rich source of user data such as IT, marketing, hospitality, and consumer sectors have a higher percentage of domains that have policies. This is true for sectors that could potentially handle sensitive data as well such as medicine, finance, human resources, and non-profit sectors. On the other, non-consumer facing sectors such as civil, mechanical and electrical as well as government and defence have the lowest percentage of websites with policies. However the *sports, media and entertainment* sector which potentially could serve as a rich source of user data has one of the lowest percentage of domains with policies, potentially indicating an issue with policy availability.

6 DISCUSSION

A user's journey to find a typical website's privacy policy would generally involve clicking on the privacy policy hyperlink on the website's landing page. However, based on our estimate, the chance that the user would find a privacy policy URL on the website is only about 34%. If the user does find the URL, we estimate that there is a 1.37% chance that the link leads to an error page. If the user is successfully directed to the privacy policy page by clicking on the link, we estimate that there is 1.23% chance that the privacy policy has no content (on an English language website). Further, if the user is on a website that is offered in multiple languages, we estimate that there is a 41.9% chance that there is a mismatch between the language in which the privacy policy is offered and the user's choice. If the user finally accesses the privacy policy without encountering any of the above failure modes, they still must spend a considerable amount of time trying to understand the privacy policy.

The General Data Protection Regulation (GDPR) is a regulation in European Union law on data protection and privacy. The GDPR has been hailed as a landmark in privacy protection law and has inspired a number of other user privacy laws around the world. According to the GDPR, entities that collect personal information from its users need to provide notice regarding what information they are collecting and why they are collecting it. *Notice and Choice* is one of primary rationales behind the GDPR and thus works only when privacy policies are readily available and easily readable. However, the lack of widespread availability of privacy policy documents online provides a basic obstacle to notice and choice. It is therefore concerning that only about 34% of websites provide a privacy policy. We acknowledge the fact that some websites that don't post a privacy policy might indeed not collect any personal information from its users. However, there is no way for the user to know whether a website collects any information or whether they are simply unable to find the policy if the company has not posted a

policy on its website. Additionally, most web servers automatically log server traffic data that includes user IP address, pages visited and time of visit. Websites might also have integration with third party services that might in-turn collect user information. It is therefore injudicious to assume that companies that have not made privacy policies available on their websites simply do not include any third-party integration and have disabled automatic user data collection features. To further complicate the problem, the nature of regulation enforcement can also pose a serious challenge since some jurisdictions (such as the European Union) rely on user complaints to investigate privacy practices [36] while others (such as the United States) only investigate when a privacy policy is at odds with real world privacy practices. We acknowledge that some websites that don't post a privacy policy might be present in jurisdictions that do not require one. However, users from parts of the world with different privacy expectations might be able to access websites that are not legally obligated to post a privacy policy. We therefore argue that posting a privacy policy should be a standard practice for any entity or organisation.

The availability of privacy policies and the ability of users to understand them are fundamental to ensuring that individuals can make informed decisions about their personal information. However, the lack of standardization and the sheer volume of privacy policies on the web can create significant barriers to access and understanding. This work contributes to ongoing efforts to promote transparency and accountability in online data privacy practices, which are critical to the continued growth and development of the digital economy. In particular, this research provides important insights into the current state of privacy policy practices on the web, including an estimate of the total number of privacy policies, and the types of organizations that are most likely to have privacy policies in place. These findings can inform efforts to develop more effective privacy policy standards and best practices, as well as to improve the accessibility and comprehensibility of existing policies for users. Additionally, this work can help policymakers and industry stakeholders better understand the challenges and opportunities associated with online data privacy practices. For example, in our study, we found that while websites that could potentially serve as a rich source of user data are likely to contain a privacy policy, websites in the *sports, media and entertainment* sector contain a unusually low percentage of websites with policies. As data privacy regulations continue to evolve and expand globally, it is essential that researchers and practitioners in the community stay up-to-date with the latest trends and best practices in this area. By contributing to this body of knowledge, this research can help support the development of more effective policies and practices that protect individual privacy rights while also facilitating innovation and growth in the digital economy. The issues regarding the availability of privacy policies is a multi-faceted problem. Thus, combined effort from regulators and researchers in multiple disciplines is required to improve trends in user data protection online.

7 CONCLUSION

Online privacy policies are crucial in safeguarding users' personal data and privacy rights. However, obtaining these policies can be a challenging task. In this study, we analyzed the online privacy policy landscape and studied the unavailability of privacy policies derived from a large dataset of company domains. At various stages of document collection and classification, we encountered a number of failures in obtaining policies and estimated frequencies of unavailability. The different failures include broken links, language inconsistencies between the document and its landing page, placeholder texts, and empty content on the privacy policy document page. We publish this novel corpus, containing 600k privacy policy corpus (with rich metadata) to the research community for further analysis.

Using a *capture-recapture technique*, we established a lower bound estimate of 2,981,047 for the total number of English language privacy policies available on the web. This result also highlights the effectiveness of multidisciplinary approaches in the field of privacy. By analyzing the distribution of privacy policies across various top-level domains (TLDs), we found that the majority of English privacy policies on the web are hosted on domains with .com, .uk, and .org TLDs. Furthermore, we demonstrated that the analysis of the distribution of privacy policies in TLDs can be leveraged to make predictions about the efficacy of privacy regulations in specific jurisdictions. Notably, our analysis revealed an unusually low number of .in domains with privacy policies in comparison to other English-speaking country-level TLDs such as .uk, .ie, .au, and .ca. We attribute this finding to the absence of national privacy regulations requiring the use of privacy policies in India. Additionally, our examination of the distribution of policies across different sectors of commerce suggests that domains within the education, sports, media, and entertainment sectors are more likely to not contain privacy policies than domains in other sectors.

Future work can explore differences and similarities of privacy policies across domains and industries while comparing instances of their failure modes. Studying failure modes based on industry regulations as well as jurisdictional regulations can further illuminate the online privacy policy landscape and therefore enable researchers and regulators to draw meaningful insight regarding privacy policies on the web.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2138131.

REFERENCES

[1] Abdel-Jaouad Aberkane, Seppe Vanden Broucke, and Geert Poels. 2022. Investigating Organizational Factors Associated with GDPR Noncompliance using Privacy Policies: A Machine Learning Approach. In *2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*. IEEE, 107–113.

[2] Ryan Amos, Gunes Acar, Elena Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2020. Privacy Policies over Time: Curation andAnalysis of a Million-Document Dataset. *arXiv preprint arXiv:2008.09159* (2020).

[3] California State Assembly. 2020. California Consumer Privacy Act. https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

[4] Michael Begon et al. 1979. *Investigating animal abundance: capture-recapture for biologists*. Edward Arnold (Publishers) Ltd.

[5] Dankmar Böhning, Irene Rocchetti, Antonello Maruotti, and Heinz Holling. 2020. Estimating the undetected infections in the Covid-19 outbreak by harnessing capture–recapture methods. *International Journal of Infectious Diseases* 97 (2020), 197–201.

[6] Hermann Brenner. 1995. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology* (1995), 42–48.

[7] Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 380–388.

[8] Lorrie Faith Cranor, Candice Hoke, Pedro Giovanni Leon, and Alyssa Phung Au. 2014. Are They Worth Reading? An In-Depth Analysis of Online Advertising Companies' Privacy Policies.

[9] Adrian Dobra and Stephen E Fienberg. 2004. How Large Is the World Wide Web? In *Web dynamics*. Springer, 23–43.

[10] Tatiana Ermakova, Benjamin Fabian, and Eleonora Babina. 2015. Readability of Privacy Policies of Healthcare Websites. *Wirtschaftsinformatik* 15 (2015).

[11] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. 2017. Large-scale readability analysis of privacy policies. In *Proceedings of the International Conference on Web Intelligence*. 18–25.

[12] Sonu Gupta, Ellen Poplavska, Nora O'Toole, Siddhant Arora, Thomas Norton, Norman Sadeh, and Shomir Wilson. 2022. Creation and Analysis of an International Corpus of Privacy Laws. *arXiv preprint arXiv:2206.14169* (2022).

[13] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *Proceedings of the 27th USENIX Conference on Security Symposium* (Baltimore, MD, USA) *(SEC'18)*. USENIX Association, USA, 531–548. https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[14] Candice Hoke, Lorrie Faith Cranor, Pedro Giovanni Leon, and Alyssa Phung Au. 2015. Are They Worth Reading? An In-Depth Analysis of Online Trackers' Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society* (2015).

[15] Madian Khabsa and C Lee Giles. 2014. The number of scholarly documents on the public web. *PloS one* 9, 5 (2014), e93949.

[16] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Technical Report. Naval Technical Training Command Millington TN Research Branch.

[17] Michael Kretschmer, Jan Pennekamp, and Klaus Wehrle. 2021. Cookie banners and privacy policies: Measuring the impact of the GDPR on the web. *ACM Transactions on the Web (TWEB)* 15, 4 (2021), 1–42.

[18] Priya C. Kumar. 2016. Privacy Policies and Their Lack of Clear Disclosure Regarding the Life Cycle of User Information. In *AAAI Fall Symposia*.

[19] Steve Lawrence and C Lee Giles. 1998. Searching the world wide web. *Science* 280, 5360 (1998), 98–100.

[20] Jianguo Lu and Dingding Li. 2010. Estimating deep web data source size by capture–recapture method. *Information retrieval* 13, 1 (2010), 70–95.

[21] Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, Jeju Island, Korea, 25–30. https://www.aclweb.org/anthology/P12-3005

[22] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 141–150.

[23] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Isjlp* 4 (2008), 543.

[24] Gabriele Meiselwitz. 2013. Readability assessment of policies and procedures of social networking sites. In *International Conference on Online Communities and Social Computing*. Springer, 67–75.

[25] Kenneth H Pollock, James D Nichols, Cavell Brownie, and James E Hines. 1990. Statistical inference for capture-recapture experiments. *Wildlife monographs* (1990), 3–97.

[26] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. 2016. Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 77–96. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/rao

[27] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4949–4959.

[28] Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, and Norman Sadeh. 2016. Automatic extraction of opt-out choices from privacy policies. In *2016 AAAI Fall Symposium Series*.

[29] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2774–2779.

[30] Mukund Srinath, Soundarya Nurani Sundareswara, C Lee Giles, and Shomir Wilson. 2021. PrivaSeer: A Privacy Policy Search Engine. In *International Conference on Web Engineering*. Springer, 286–301.

[31] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 6829–6839.

[32] Peter Story, Sebastian Zimmeck, Abhilasha Ravichander, Daniel Smullen, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Natural language processing for mobile app privacy compliance. In *AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*.

[33] Soundarya Sundareswara, Shomir Wilson, Mukund Srinath, and Lee Giles. 2020. Privacy not found: a study of the availability of privacy policies on the web.

[34] Soundarya Nurani Sundareswara, Mukund Srinath, Shomir Wilson, and C. Lee Giles. 2021. A Large-Scale Exploration of Terms of Service Documents on the Web. In *Proceedings of the 21st ACM Symposium on Document Engineering* (Limerick, Ireland) *(DocEng '21)*. Association for Computing Machinery, New York, NY, USA, Article 21, 4 pages. https://doi.org/10.1145/3469096.3474940

[35] Ali Sunyaev, Tobias Dehling, Patrick Taylor, and Kenneth Mandl. 2014. Availability and Quality of Mobile Health App Privacy Policies. *Journal of the American Medical Informatics Association* (08 2014), 1–4. https://doi.org/10.1136/amiajnl-2013-002605

[36] Factsheet 5: European Data Protection Supervisor. 2018. What to expect when we inspect. (2018). https://edps.europa.eu/sites/edp/files/publication/18-11-21_factsheet_inspections_en.pdf

[37] Verisign. [n. d.]. *VERISIGN Q1 2020 DOMAIN NAME INDUSTRY BRIEF*. https://blog.verisign.com/domain-names/verisign-q1-2020-domain-name-industry-brief-internet-grows-to-366-8-million-domain-name-registrations-in-the-first-quarter-of-2020/

[38] Rhiannon Weaver and M Patrick Collins. 2007. Fishing for phishes: Applying capture-recapture methods to estimate phishing populations. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*. 14–25.

[39] Wikipedia contributors. 2020. *Lorem ipsum — Wikipedia, The Free Encyclopedia*. [Online; accessed 12-May-2020].