The University of Texas at Austin
Center for Identity

PrivacyCheck's Machine Learning to Digest PrivacyPolicies: Competitor Analysis and Usage Patterns

Razieh Nokhbeh Zaeem
Safa Anya
Alex Issa
Jake Nimergood
Isabelle Rogers
Vinay Shah
Ayush Srivastava
K. Suzanne Barber

UTCID Report #20-10

June 2020

PrivacyCheck's Machine Learning to Digest Privacy Policies: Competitor Analysis and Usage Patterns

1st Razieh Nokhbeh Zaeem
The Center for Identity
The University of Texas at Austin
razieh@identity.utexas.edu

2nd Safa Anya
Electrical and Computer Engineering Dept.
The University of Texas at Austin
safaianya@gmail.com

3rd Alex Issa
Electrical and Computer Engineering Dept.
The University of Texas at Austin
alex.issa32@utexas.edu

4th Jake Nimergood
Electrical and Computer Engineering Dept.
The University of Texas at Austin
jakenimergood@gmail.com

5th Isabelle Rogers
Electrical and Computer Engineering Dept.
The University of Texas at Austin
isabelle.rogers97@gmail.com

6th Vinay Shah
Electrical and Computer Engineering Dept.
The University of Texas at Austin
vinayshah@utexas.edu

7th Ayush Srivastava
Electrical and Computer Engineering Dept.
The University of Texas at Austin
ayushsriv@utexas.edu

8th K. Suzanne Barber
The Center for Identity
The University of Texas at Austin
sbarber@identity.utexas.edu

***Abstract*—Online privacy policies are lengthy and hard to comprehend. To address this problem, researchers have utilized machine learning (ML) to devise tools that automatically summarize online privacy policies for web users. One such tool is our free and publicly available browser extension, PrivacyCheck. In this paper, we enhance PrivacyCheck by adding a competitor analysis component—a part of PrivacyCheck that recommends other organizations in the same market sector with better privacy policies. We also monitored the usage patterns of about a thousand actual PrivacyCheck users, the first work to track the usage and traffic of an ML-based privacy analysis tool. Results show: (1) there is a good number of privacy policy URLs checked repeatedly by the user base; (2) the users are particularly interested in privacy policies of software services; and (3) PrivacyCheck increased the number of times a user consults privacy policies by 80%. Our work demonstrates the potential of ML-based privacy analysis tools and also sheds light on how these tools are used in practice to give users actionable knowledge they can use to pro-actively protect their privacy.**

***Index Terms*—privacy policy, machine learning, PrivacyCheck, usable privacy, browser extension, privacy enhancing technologies, PET, competitor analysis**

I. Introduction

Many websites collect, share, and use their users' Personally Identifiable Information (PII)—"any information relating to an identified or identifiable natural person" [1]. An online privacy policy is a legal document that informs the users about the PII collection, sharing, and usage practices of a website. Many regulatory bodies around the globe have long enforced requirements on posting privacy policies online. Over the past few decades, however, the collection, sharing, and usage of users' PII have risen to a major privacy concern over the Internet, so much so that newer laws have gone into effect to protect user privacy. Prominent examples of such laws are the General Data Protection Regulation (GDPR)[1] in the European Union and the California Consumer Privacy Act (CCPA)[2] in the United States.

Research has shown, time and again [2]–[5], that privacy policies are simply too long and hard to comprehend for their intended users, and therefore users rarely take the time and effort to read them. To address the poor readability of privacy policies, researchers have developed tools that leverage Machine Learning (ML) to automatically summarize privacy policies. Among these tools, few are made publicly available.

One of these tools that leverages ML to summarize an online privacy policy is our own free and publicly available browser extension, PrivacyCheck [6]. When the user navigates to a privacy policy in the browser, he/she can run PrivacyCheck to recap the privacy policy with ML. The first version of PrivacyCheck used to summarize privacy policies based on ten *User Control* privacy questions that were rooted in the work of the Organization for Economic Cooperation and Development [7], and the Federal Trade Commission (FTC) Fair Information Practices (FIP) [8]. Recently, we presented the second version of PrivacyCheck (briefly covered in a short tool paper [9]). This new version adds ten new *GDPR* questions [9], [10]. The second version of PrivacyCheck inherited about a thousand users from its first version, all of whom received an automatic update to the new version.

In this paper, we make the following contributions over our previous work [6], [9]:

1) We explain the technical development of a new component of PrivacyCheck–the **Competitor Analysis Tool**

[1]https://gdpr-info.eu
[2]https://oag.ca.gov/privacy/ccpa

(CAT)–that suggests the best competitors (with respect to user privacy) in the same market sector as of the policy under investigation. Our CAT maintains a database of privacy policy summaries in each market sector. This database (which does not contain any PII from PrivacyCheck users) enables CAT to suggest competitors with better privacy policies. We initially populate this database with over 4,000 privacy policies across market sectors.

2) As users run PrivacyCheck, our CAT gradually improves its database. Equally importantly, such database can serve as a treasure trove of insights about what privacy policies are analyzed by typical users. **We study this database and the usage traffic of PrivacyCheck to understand the behavior of the users of an ML-based privacy analysis tool.**

We observe that PrivacyCheck users have a tendency to check the same privacy policies, presumably the most commonly used services or the most important sources of privacy concerns. There is, generally, a good number of URLs investigated by the PrivacyCheck user base more than once: among 534 calls to PrivacyCheck, only 366 (68%) were on unique URLs. The average rate of new policies added to the CAT database by the entire user base of PrivacyCheck was about 2 per day over the first three months.

We find it fascinating that the user base of PrivacyCheck was disproportionately interested in running it on a variety of software service privacy policies (online social networks, large software companies, and predominantly smaller software companies) versus any other market sector/category. We think that the reason is the sheer amount of information such services can collect from their users.

The pre-populating phase of the CAT database with over 4,000 privacy policies was a necessary first step to provide meaningful CAT results upon PrivacyCheck release. The URLs added to CAT in this phase, however, were a random sample of crawling the web for privacy policies. Monitoring the usage patterns of PrivacyCheck, we find that few of these URLs (only 16 out of 4,273) are of interest to typical PrivacyCheck users.

Finally, we report that ML-based privacy analysis tools have the potential to increase the number of times a typical user consults privacy policies. PrivacyCheck improved the number of times a user investigates a privacy policy by 80%, from 1% of users to 1.8%. This improvement is tangible, but is still far from ideal.

II. Background: PrivacyCheck Browser Extension

PrivacyCheck is available online[3] as a Google Chrome browser extension. Figure 1 shows a screen-shot of its new version. When the user navigates to an online privacy policy and runs PrivacyCheck, the client side browser extension sends the URL of the policy to the PrivacyCheck server, running on Amazon Web Services (AWS). The server executes (1) the machine learning classification models to answer ten User Control and ten GDPR questions and (2) the competitor analysis. The server then sends the results back to the browser extension to display for the user.

The PrivacyCheck server utilizes its ML models to analyze the privacy policy text to automatically answer ten *User Control* questions and ten *GDPR* questions (e.g., Figure 2). Combining the answers to each set of these ten questions generates an overall score for the policy, one with respect to User Control and one pertaining to GDPR (as shown in Figure 1). The User Control questions were developed as a part of the first version of PrivacyCheck [6]. The GDPR questions and the machine learning corpora used to train their models were designed in our previous work [10]. In that work, we envisioned how future PrivacyCheck advances can support the new GDPR questions to further identify GDPR compliances [9]. For completeness, Table I lists the User Control and GDPR questions.

In addition to the ML models, the PrivacyCheck server implements the CAT functionality. The CAT reports top three competitors and their scores, i.e., top three privacy policies in the same sector that have received the highest scores from PrivacyCheck. The CAT also puts the privacy score in context by reporting the mean (average) privacy score in the corresponding sector. For either of User Control and GDPR, CAT shows these results separately. For instance, Figure 3 shows the CAT panel for GDPR. We initially populated the CAT database with over 4,000 privacy policy URLs, their market sectors, User Control/GDPR scores, and answers to the twenty questions (automatically answered by PrivacyCheck ML models). Furthermore, as PrivacyCheck users run it on various privacy policies over time, the CAT collects more URLs and their market sectors and scores. We elaborate on the CAT in the next section.

III. The Competitor Analysis Tool

While the previous versions of PrivacyCheck merely aimed to *educate* and *inform* users on data practices, our competitor analysis tool *empowers* users to choose services with better privacy policies. After scoring a policy, the user can view a list of the three companies with the highest GDPR or User Control scores within that same market sector. The CAT also displays a graph that shows how that company's score compares to the mean score in the company's market sector.

In order for the CAT to recommend competitor websites with better User Control or GDPR privacy policy scores, we need to automatically

A) Detect the market sector/category of a privacy policy,

B) Calculate the User Control/GDPR scores for the policy,

C) Have a pre-populated database of URLs across categories/sectors, their market sectors, and their scores, but collect more URLs as we go,

D) Search (with high efficiency) in the database to find competitors and average scores in the market sector.

[3]https://chrome.google.com/webstore/detail/privacycheck/poobeppenopkcbjejfjenbiepifcbclg?hl=en-US

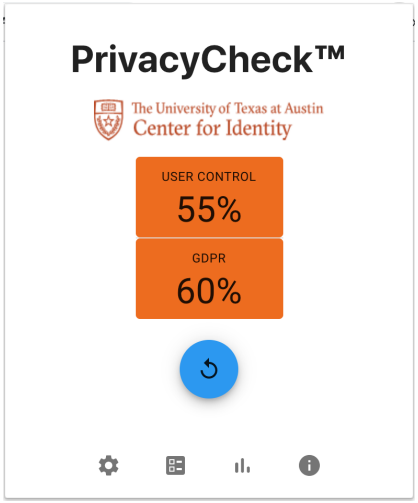

Fig. 1: PrivacyCheck: User Control and GDPR scores of a sample policy.

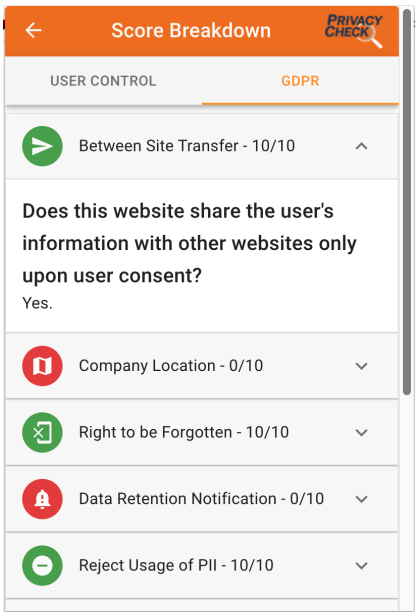

Fig. 2: GDPR score breakdown panel.

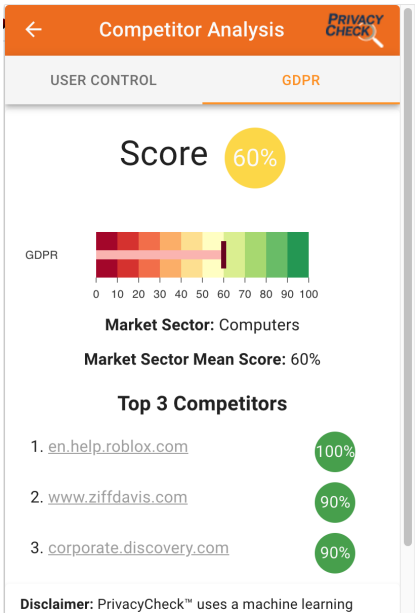

Fig. 3: GDPR CAT panel.

TABLE I: PrivacyCheck complete set of questions.

	User Control
1	How well does this website protect your email address?
2	How well does this website protect your credit card information and address?
3	How well does this website handle your social security number?
4	Does this website use or share your PII for marketing purposes?
5	Does this website track or share your location?
6	Does this website collect PII from children under 13?
7	Does this website share your information with law enforcement?
8	Does this website notify or allow you to opt-out after changing their privacy policy?
9	Does this website allow you to edit or delete your information from its records?
10	Does this website collect or share aggregated data related to your identity or behavior?
	GDPR
1	Does this website share the user's information with other websites only upon user consent?
2	Does this website disclose where the company is based/user's PII will be processed & transferred?
3	Does this website support the right to be forgotten?
4	If they retain PII for legal purposes after the user's request to be forgotten, will they inform the user?
5	Does this website allow the user the ability to reject usage of user's PII?
6	Does this website restrict the use of PII of children under the age of 16?
7	Does this website advise the user that their data is encrypted even while at rest?
8	Does this website ask for the user's informed consent to perform data processing?
9	Does this website implement all of the principles of data protection by design and by default?
10	Does this website notify the user of security breaches without undue delay?

A. Detecting the Market Sector of a Privacy Policy

For the CAT, we created an ML classifier that determines the market sector using the URL of a privacy policy. Inspired by the work of Shawon et al. [11], we utilized the DMOZ dataset [12] of 1.5 million URLs and trained a classifier that categorizes a given URL into one of the 15 market sectors/categories. These categories, as explained in [11] and sorted alphabetically, are: (1) adult, (2) arts, (3) business, (4) computers, (5) games, (6) health, (7) home, (8) kids, (9) news, (10) recreation, (11) reference, (12) science, (13) shopping, (14) society, and (15) sports.

We first applied tf-idf pre-processing to the URL itself, and then trained a Logistic Regression model on the DMOZ dataset. This model varies from the original work [11] in that it switched from Multinomial Naive Bayes to Logistic Regression. We initially trained a Multinomial Naive Bayes model but found its size (2.2 GB) prohibitively large for our application, because the AWS API that loads the model cannot do so within the AWS API Gateway's 30s timeout window. Instead, we trained a Logistic Regression model (200 MB) that takes an average of about 8s to load and output a classification. Changing the Multinomial Naive Bayes classifier to Logistic Regression, however, did involve some compromise in terms of classification accuracy. To measure accuracy of classification,

we split the DMOZ dataset into training and validation sets following normal data science conventions of 70% training and 30% validation data. Our implementation of the Multinomial Naive Bayes model attained an average of 87% accuracy across categories. Our Logistic Regression achieved an average accuracy of 56% across the 15 categories. A random classifier would be only 7% accurate. Table II shows the accuracy measures of our 200 MB Logistic Regression URL classifier, which we used in the implementation.

TABLE II: Logistic Regression URL classification measures.

Category	Precision	Recall	F1-Score	Support
Adult	0.92	0.34	0.49	9933
Arts	0.41	0.55	0.47	75204
Business	0.30	0.85	0.44	71280
Computers	0.76	0.20	0.31	34854
Games	0.78	0.44	0.56	16523
Health	0.81	0.17	0.28	17512
Home	0.89	0.34	0.49	7856
Kids	0.53	0.23	0.32	13368
News	0.57	0.07	0.12	2760
Recreation	0.61	0.08	0.14	30803
Reference	0.50	0.48	0.49	16812
Science	0.55	0.36	0.44	32258
Shopping	0.64	0.02	0.03	27826
Society	0.49	0.47	0.48	72822
Sports	0.92	0.29	0.45	30083
Accuracy			0.42	459894
Macro Avg.	0.65	0.32	0.37	459894
Weighted Avg.	0.56	0.42	0.39	459894

B. Calculating User Control/GDPR Scores

Our ML classification models of PrivacyCheck answer the User Control and GDPR questions and then calculate the scores. Our previous work details the User Control models [6] and the GDPR models [9], [10].

C. The Database of Policies, Market Sectors, and Scores

We created the database for the CAT as a DynamoDB database on AWS. AWS DynamoDB is a NoSQL database hosted in AWS that PrivacyCheck utilizes to store and retrieve data about privacy policies efficiently. Entries are stored using the privacy policy URL as the primary key, and the other fields include the company's domain URL, the market sector, GDPR and User Control scores, and the date that the entry was made.

We first populated the CAT database leveraging the DMOZ dataset [12]. DMOZ (also known as the Open Directory Project) is a huge and comprehensive manually edited directory of the web, which contains over 1.5 million URLs in 15 categories. We selected a 1% random sample of the 1,562,978 URLs of DMOZ, downloaded the page of each URL, and found all the links on the page that point to what PrivacyCheck considers a privacy policy URL[4]. Parsing this sample produced 9,579 privacy policy links but only 4,273 links (45%) were unique. We believe that such high percentage of shared privacy policies is due to shared parent companies and widespread use of template policies. Table III displays the number of unique privacy links based on the 1% sampling of DMOZ, divided by categories automatically assigned by PrivacyCheck's Logistic Regression URL classifier. We used these privacy policies to populate the CAT database before beta testing and release.

A closer look at Table III reveals that there are no policies in the News category. The reason is this category has the second lowest recall and the lowest support in Table II, and hence our URL classifier is failing to detect policies in the News category. Overall, the distribution of URLs in the categories of this table is different than the distribution of URLs in DMOZ, which we assume to be somewhat representative of the entire web. In particular, there are too many privacy policy URLs in the Computers and Health categories of this table, when compared with DMOZ [11]. Manual investigation of privacy policy URLs and the measures reported in Table II reveal that, different from the News category, the high number of URLs in these categories is not due to classification error. A possible reason could be that there are relatively more web pages in Computers and Health categories that do have privacy policies, therefore creating more policies in these categories.

As a part of future work, we plan to add more privacy policy URLs to the CAT database, through more than 1% of DMOZ. However, the use of more URLs incurs AWS expenses that limited our first collection of URLs for CAT. In addition, as we point out in Section IV-A, pre-populating the CAT database with a random sample is not necessarily optimal. Consequently, we currently rely on users to populate CAT with privacy policies that are of more interest to users versus a random sample obtained from crawling the web.

D. Efficient Search in the Database

In the CAT database, the entries are kept unique by their privacy policy URL, which acts as the primary key. However, we also need the ability to search the database for entries with a particular market sector in order to implement the CAT. Doing this search using "scan" operations would be very costly and require lots of computation time. Instead, we leveraged the Global Secondary Index (GSI) feature of DynamoDB, which essentially creates a smaller set of tables within the database that took fields from the existing database but could have a separate primary key. Two GSIs were created: one sorted by market sector of the companies scored using GDPR, and a similar one for User Control. The GSIs allow quick searches on the database for companies in the market sector in question, to find the top three competitors and the mean score in the market sector.

IV. Lessons Learned from PrivacyCheck Usage

In this section we investigate the content of the CAT database and also the traffic of PrivacyCheck to obtain insights into the usage patterns of an ML-based privacy analysis tool. As of this writing, the second version of PrivacyCheck has 911 users, most of which it inherited from the first version. PrivacyCheck has a rating of 4.3 out of 5, based on 12 reviews.

[4]PrivacyCheck leverages a regular expression to determine whether a URL belongs to a privacy policy [9].

TABLE III: Number of URLs in the pre-populated database.

Phase	Date	Total	Adult	Arts	Business	Computers	Games	Health	Home	Kids	News	Recreation	Reference	Science	Shopping	Society	Sports
Pre-populating (2 days)	5/01/2020 to 5/02/2020	4273	27	269	362	1470	11	354	12	46	0	25	167	222	33	1217	58

A. Lessons Learned from the CAT Database

The new version of PrivacyCheck was released to the public on May 24, 2020. We divide the items in the CAT database into three groups:

1) Policies added or updated during the beta testing of PrivacyCheck by a class of undergraduate students at the University of Texas at Austin—May 3 to May 24, 2020.
2) Policies added/updated by the user base of PrivacyCheck during the first month after the release—May 25 to June 24, 2020.
3) Policies added/updated by its user base during the second and third months after the release—June 25 to August 24, 2020.

Table IV displays the number of policies added/updated in each phase by market sector/category. We observe that users of PrivacyCheck have been trying it on new policies. Note that if PrivacyCheck is run on a policy that already exists in its database, it merely updates the corresponding entry. In the first month after release, an average of 2.00 new policies per day were added to the CAT database. In the second and third months, an average of 2.05 new policies per day were added. We add that only 16 policies pre-populated in the CAT database where ever called by the beta testers and actual users (11 in Computers, 3 in Society, 1 in Science, and 1 in Sports). These results imply that the random pre-population of the CAT database is a necessary first step to provide meaningful CAT results, but does not ultimately collect the URLs that are frequently analyzed by actual users.

Considering Table IV, one market sector/category is by far more populated than others: Computers. Not only does this category appear more often in the beta testing phase, but also in the URLs added by the users during the first three months. Interestingly, according to Shawon et al. [11], this skewed distribution of categories in the CAT database is not a result of an underlying skew in the web. As they demonstrate, the Arts, Society, and Business categories are the largest categories in the DMOZ dataset and each have at least twice as many URLs as the Computers category. (Also see the support column of 30% of DMOZ shown in Table II.) Assuming that the 1.5 million URLs in DMOZ are representative of all the 200 million websites on the Internet[5], the Computers category is not the most populated on the web. Furthermore, the 76% classification precision of the Computers category (Table II) makes it unlikely that those URLs are assigned to this category by mistake. In fact, the low recall of 20% for this category may suggest that there are even more URLs that should have been assigned to Computers but were not.

We manually investigated the *Computers* URLs added or updated by the actual user base of PrivacyCheck (excluding the URLs only added or updated in beta testing). In these 187 URLs, we find that 72% were correctly categorized as computer-related, a precision that is close to the 76% precision reported in Table II. We further observe that 9% belong to online social networks (e.g., Twitter and Facebook), 17% are large international software companies (e.g., Google and IBM) and 35% are smaller software companies (covering a wide range including services on security and privacy, online proctoring, online communication, productivity software, VPN, etc.).

From monitoring the CAT database over the first three months after its release to the public, we see a disproportionate attention from the users of PrivacyCheck paid to online software services. While many researchers have quantified the user interest, or lack thereof, in privacy policies, we are the first to find that the user base of PrivacyCheck (however small compared to the total number of web users) was very interested in running it on a variety of software service privacy policies. The presence of online social media privacy policies and huge international software companies in the list is more or less expected. It is the smaller software services, however, that we also find frequently analyzed. A possible reason is the amount of information software services collect, from conversations in online communication software to one's entire web traffic in a VPN service. Other types of information collected from users in other market sectors (for example when shopping online for clothes) seem pale in comparison. A paramount future work for us is to study usage patterns of PrivacyCheck over a longer period of time, with built in ML-based categorization tuned to more granular sub-sectors of the Computers sector.

B. Lessons Learned from PrivacyCheck Traffic

Figure 4 depicts the number of calls to the PrivacyCheck AWS server over the first three months (from May 24, 2020 to August 24, 2020), as reported by AWS. The number of calls are accumulated over the week ending in a given date, e.g., in the week that includes 06/21, users executed PrivacyCheck on 151 URLs, 41 of which were seen for the first time. Consequently, PrivacyCheck has already been executed on the other 110 policies during the beta testing and after release.

From Figure 4, we make a number of conclusions:

1) Many URLs on which users run PrivacyCheck are already investigated by others, even over the relatively

[5]https://news.netcraft.com/archives/category/web-server-survey

TABLE IV: Number of PrivacyCheck executions and CAT database entries.

Phase	Date	Executions	CAT Entries Added/Updated.															
			Total	Adult	Arts	Business	Computers	Games	Health	Home	Kids	News	Recreation	Reference	Science	Shopping	Society	Sports
Beta Testing (22 days)	5/03/2020 to 5/24/2020	711	179	2	4	12	113	1	3	0	0	0	0	1	2	2	37	2
Month 1 (31 days)	5/25/2020 to 6/24/2020	114	62	0	0	2	46	0	0	0	0	0	0	0	0	0	14	0
Months 2 & 3 (61 days)	6/25/2020 to 8/24/2020	364	125	0	3	7	87	0	2	0	0	0	0	0	1	1	24	0

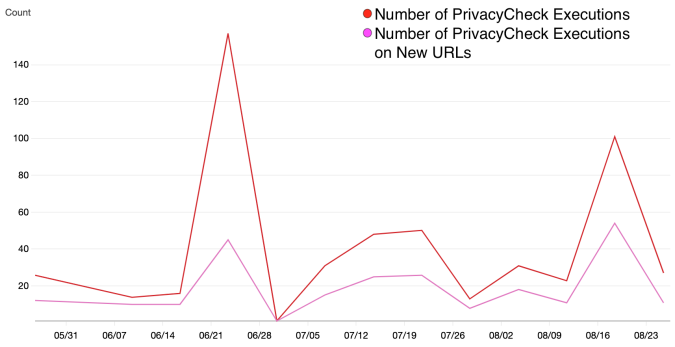

Fig. 4: Weekly PrivacyCheck execution statistics in the first three months since the release (May 24 to August 24, 2020).

short time period of three months. This observation confirms our prior assumption that collecting privacy scores in a database and fetching results instead of recalculating them improves the performance of PrivacyCheck [9]. The pre-populating phase, however, does not play a big role with its 16 shared URLs between pre-population and testing/actual use. As a result, the collection of URLs from users is a more viable option compared to pre-populating the database, at least when the goal is to improve PrivacyCheck performance or put its score in the context of *widely-used* privacy policies.

2) An estimate [3] counts the number of new sites each user visits per year at 119. With 911 users of PrivacyCheck and over the period of three months (0.25 of a year), we approximate that the user base of PrivacyCheck has encountered just over 27,000 new websites over these three months. With 478 (114 + 364 from Table IV) total calls to run PrivacyCheck, the user base of PrivacyCheck is running it on 1.8% of the new websites they see. Indeed, PrivacyCheck has improved the frequency of assessing privacy policies, when we compare to the baseline server side observation that only 1% or less of users click on a website's privacy policy [13] in the absence of ML-based tools. While the 80% improvement is promising, the 1.8% of users checking out privacy policies is still far from ideal.

V. Threats to Validity

The most important threat to the external validity of our results is a possible sampling bias. We measured the percentage of times a typical PrivacyCheck user investigates a privacy policy at 1.8%, and compared it with 1%—the server side measurement of the percentage of users who click on the privacy policy link [13]. A potential sampling bias might exist, in that the users who installed PrivacyCheck might already have an interest in privacy and not represent the general population.

VI. Related Work

Researchers have studied, for many years, when and how users read (or ignore) privacy policies. Most of such studies focus on the percentage of users who do read these policies and poor policy readability. To circumvent the readability issue with these policies, many researchers have utilized various machine learning methodologies to automatically summarize privacy policies.

In this section, we cover related work with respect to our competitor analysis tool as well as the study of the usage patterns of PrivacyCheck. We are unaware of any work that provides usage statistics of ML-based privacy tools. Nonetheless, we cover the most prominent work on when and how users read privacy policies. Finally, most of the ML-based privacy analysis tools are are to *inform* users and our CAT that directly provides actionable advice is rare and novel. We briefly cover closely related work, with an emphasis on available tools.

A. Reading Patterns of Privacy Policies

The topic of user interest, or lack thereof, in privacy policies has been on the minds of many researchers. McDonald and Cranor [3] noted that if users were to read the privacy policy for each site they visit just once a year, they would need to spend over 200 hours doing so. In fact, less than half of website users claim to have *ever* read a privacy policy [14]. Studies that used self-reported data from users found that only 4.5% claim to always read them [15] and more reliable server side observation of websites reveals that only 1% or less of users click on a website's privacy policy [13]. More recent work [5] demonstrated that three out of four users completely ignore privacy policies. Other users skim through policies that take 29 to 32 minutes to read in less than two minutes.

The fact that most users do not read privacy policies might be attributed to the privacy policies' poor readability [16]. A study of the readability of privacy policies showed that the average privacy policy required two years of college level education to comprehend [2], [17].

B. CAT Related Work

One of the ways to address the poor readability of privacy policies is to utilize ML-based tools that automatically summarize privacy policies. To our knowledge, Privee [18] is the first tool to automatically analyze privacy policies. Building on the crowd sourcing privacy analysis framework ToS;DR [19], Privee uses machine learning to classify privacy policies that are not already rated in ToS;DR. As opposed to PrivacyCheck, Privee predates the GDPR and as a result does not address it.

Polisis [20] is a browser extension that utilizes deep learning to evaluate the PII collected and shared according to a privacy policy. Pribots [21] is a chat-bot from the same authors that answers free-form questions about privacy policies. Close to our CAT but different is their PoliCompare tool[6] that explicitly asks for the URL of up to ten privacy policies and compares them together.

The Usable Privacy Project takes advantage of natural language processing and machine learning to semi-automatically annotate privacy policies. This project annotates [22], [23] OPP-115 (a corpus of 115 policies with attributes and data practices), which is the corpus that Polisis uses. PrivacyCheck is distinct from these projects—particularly with its novel competitor analysis.

MAPS [24] analyzes privacy policies of more than one million mobile applications. PolicyLint [25] is a natural language processing tool that identifies potential contradictions that may arise inside the same privacy policy. PrivacyGuide [26], is a machine learning and natural language processing tool inspired by the GDPR. However, PrivacyGuide is not publicly available as a tool.

At the Center for Identity at the University of Texas at Austin, we target many aspects of identity management and privacy [27]–[31]. We developed PrivacyCheck [6], [9] and used it to study privacy policies across industries [32] and to quantify the effect of the GDPR on the landscape of privacy policies [10]. To the best of our knowledge, we are the first to study usage patterns of ML-based privacy analysis tools.

VII. Conclusion and Future Work

We detailed the technical development of the Competitor Analysis Tool (CAT), the newest addition to our ML-based privacy analysis browser extension, PrivacyCheck. We demonstrated how we initially populated the database of the CAT with the privacy policies of a random 1% sample of the widely used DMOZ web directory. We further monitored the growth of the CAT database and the usage patterns of PrivacyCheck over the first three months after release, with about one thousand users who had PrivacyCheck installed on their web browsers.

We found that 32% of the URLs on which PrivacyCheck was used to evaluate a privacy policy were repeated, at least once, during the first three months. The random URLs initially populating the CAT database from DMOZ privacy policies were not of much interest to PrivacyCheck users. Nonetheless, the pre-populating of the CAT database was necessary so that PrivacyCheck–upon its release–would have a baseline in each market sector for competitor analysis.

We found PrivacyCheck users were distinctly interested in privacy policies of the Computers sector, including online social networks (e.g., Facebook and Twitter) and big software corporations (e.g., Google and IBM). Even more interesting, PrivacyCheck users used the Competitor Analysis Tool on a longer list of lesser known software services.

Finally, we found that a typical PrivacyCheck user investigates 1.8% of all the new websites he/she visits in a year. Previous literature estimates a typical (non PrivacyCheck) web user investigates 1% of the privacy policies of new websites visited. We concluded that PrivacyCheck has the potential to increase the number of times a user consults a privacy policies and, consequently, increase their knowledge about an organizations privacy commitments.

A paramount future work for us is to study usage patterns of PrivacyCheck over a longer period of time, with built in ML-based categorization tuned to more granular sub-sectors (possibly using the subcategories of DMOZ), specifically for the Computers sector. We will also consider adding more privacy policies to the CAT database, similar to the above pre-populating of this database.

This research studied ML-based privacy policy analysis tools in a different light. Instead of merely showing how faithful our summarization results are to the actual privacy policy text (which we have done in previous work [6], [9], [10], [32]), we introduced and implemented a practical Competitor Analysis Tool in PrivacyCheck to empower users, and explored how actual users leverage ML-based privacy analysis tools in practice. Our work paves the way for tools that not only inform but empower by understanding the actual practices of real web users and by giving them actionable knowledge to use disclosures in privacy policies to select those organizations that best protect and respect their privacy.

[6]https://pribot.org/polisis/compare

Acknowledgment

We thank the Center for Identity Partners (http://identity.utexas.edu/strategic-partners) for their contributions to this research effort.

References

[1] E. Union, "European union law." [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1566668063189&uri=CELEX:32016R0679

[2] M. A. Graber, D. M. D Alessandro, and J. Johnson-West, "Reading level of privacy policies on internet health web sites," *Journal of Family Practice*, vol. 51, no. 7, pp. 642–642, 2002.

[3] A. M. McDonald and L. F. Cranor, "the cost of reading privacy policies," *I/S: A Journal of Law and Policy for the Information Society*, vol. 4, p. 543, 2008.

[4] B. Fabian, T. Ermakova, and T. Lentz, "Large-scale readability analysis of privacy policies," in *Proceedings of the International Conference on Web Intelligence*, 2017, pp. 18–25.

[5] J. A. Obar and A. Oeldorf-Hirsch, "The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services," *Information, Communication & Society*, vol. 23, no. 1, pp. 128–147, 2020.

[6] R. N. Zaeem, R. L. German, and K. S. Barber, "Privacycheck: Automatic summarization of privacy policies using data mining," *ACM Transactions on Internet Technology (TOIT)*, vol. 18, no. 4, p. 53, 2018.

[7] H. Regard, "Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data," 1980.

[8] FTC, "Privacy online: Fair information practices in the electronic marketplace: A federal trade commission report to congress," 2000. [Online]. Available: https://tinyurl.com/lav5ndf

[9] R. N. Zaeem, S. Anya, A. Issa, J. Nimergood, I. Rogers, V. Shah, A. Srivastava, and K. S. Barber, "PrivacyCheck v2: A tool that recaps privacy policies for you," in *29th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2020, to appear.

[10] R. N. Zaeem and K. S. Barber, "The effect of the GDPR on privacy policies: Recent progress and future promise," *ACM Transactions on Management of Information Systems*, 2020.

[11] A. Shawon, S. T. Zuhori, F. Mahmud, and M. J.-U. Rahman, "Website classification using word based multiple n-gram models and random search oriented feature parameters," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*. IEEE, 2018, pp. 1–6.

[12] DMOZ, "Open directory project," 2020. [Online]. Available: https://dmoz-odp.org

[13] R. Kohavi, "Mining e-commerce data: the good, the bad, and the ugly," in *International conference on Knowledge discovery and data mining*. ACM, 2001, pp. 8–13.

[14] D. B. Meinert, D. K. Peterson, J. R. Criswell, and M. D. Crossland, "Privacy policy statements and consumer willingness to provide personal information," *Journal of Electronic Commerce in Organizations*, vol. 4, no. 1, p. 1, 2006.

[15] G. R. Milne and M. J. Culnan, "Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices," *Journal of Interactive Marketing*, vol. 18, no. 3, pp. 15–29, 2004.

[16] T. Ermakova, A. Baumann, B. Fabian, and H. Krasnova, "Privacy policies and users' trust: Does readability matter?" in *20th Americas Conference on Information Systems (AMCIS)*, 2014.

[17] G. R. Milne, M. J. Culnan, and H. Greene, "A longitudinal assessment of online privacy notice readability," *Journal of Public Policy & Marketing*, vol. 25, no. 2, pp. 238–249, 2006.

[18] S. Zimmeck and S. M. Bellovin, "Privee: An architecture for automatically analyzing web privacy policies," in *23rd USENIX Security Symposium*. San Diego, CA: USENIX Association, Aug 2014, pp. 1–16.

[19] ToS;DR, "Terms of service; didn't read," 2012. [Online]. Available: https://tosdr.org

[20] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *27th USENIX Security Symposium*, 2018, pp. 531–548.

[21] H. Harkous, K. Fawaz, K. G. Shin, and K. Aberer, "Pribots: Conversational privacy with chatbots," in *Twelfth Symposium on Usable Privacy and Security (SOUPS) 2016*, 2016.

[22] S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, N. A. Smith, and F. Liu, "Crowdsourcing annotations for websites' privacy policies: Can it really work?" in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 133–143.

[23] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell *et al.*, "The creation and analysis of a website privacy policy corpus," in *Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1330–13 340.

[24] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. Reidenberg, N. C. Russell, and N. Sadeh, "Maps: Scaling privacy compliance analysis to a million apps," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 3, pp. 66–86, 2019.

[25] B. Andow, S. Y. Mahmud, W. Wang, J. Whitaker, W. Enck, B. Reaves, K. Singh, and T. Xie, "Policylint: investigating internal privacy policy contradictions on Google play," in *28th USENIX Security Symposium*, 2019, pp. 585–602.

[26] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna, "Privacyguide: towards an implementation of the EU GDPR on internet privacy policy evaluation," in *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. ACM, 2018, pp. 15–21.

[27] R. N. Zaeem, M. Manoharan, Y. Yang, and K. S. Barber, "Modeling and analysis of identity threat behaviors through text mining of identity theft stories," *Computers & Security*, vol. 65, pp. 50–63, 2017.

[28] R. N. Zaeem, S. Budalakoti, K. S. Barber, M. Rasheed, and C. Bajaj, "Predicting and explaining identity risk, exposure and cost using the ecosystem of identity attributes," in *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2016, pp. 1–8.

[29] R. N. Zaeem, M. Manoharan, and K. S. Barber, "Risk kit: Highlighting vulnerable identity assets for specific age groups," in *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 2016, pp. 32–38.

[30] J. Zaiss, R. Nokhbeh Zaeem, and K. S. Barber, "Identity threat assessment and prediction," *Journal of Consumer Affairs*, vol. 53, no. 1, pp. 58–70, 2019.

[31] R. Rana, R. N. Zaeem, and K. S. Barber, "An assessment of blockchain identity solutions: Minimizing risk and liability of authentication," in *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2019, pp. 26–33.

[32] R. N. Zaeem and K. S. Barber, "A study of web privacy policies across industries," *Journal of Information Privacy and Security*, vol. 13, no. 4, pp. 169–185, 2017.

WWW.IDENTITY.UTEXAS.EDU