The University of Texas at Austin
Center for Identity

PrivacyCheck v3: Empowering Users with Higher-Level Understanding of Privacy Policies

Razieh Nokhbeh Zaeem
Ahmad Ahbab
Josh Bestor
Hussam H. Djadi
Sunny Kharel
Victor Lai
Nick Wang
K. Suzanne Barber

UTCID Report #21-03

August 2021

PrivacyCheck v3: Empowering Users with Higher-Level Understanding of Privacy Policies

Razieh Nokhbeh Zaeem
Ahmad Ahbab*
Josh Bestor*
Hussam H. Djadi*
Sunny Kharel*
Victor Lai*
Nick Wang*
K. Suzanne Barber
University of Texas at Austin
Austin, Texas, USA
nokhbeh,amadraccoon,joshbestor,djadih,sunnykharel,laivictor2718,nickwang3@utexas.edu,sbarber@identity.utexas.edu

ABSTRACT

Privacy policies are lengthy and hard to read, yet are profoundly important as they communicate the practices of an organization pertaining to user data privacy. Privacy Enhancing Technologies, or PETs, seek to inform users by summarizing these privacy policies. Efforts in the research and development of such PETs, however, have largely been limited to tools that recap the policy or visualize it. We present the next generation of our research and publicly available tool, PrivacyCheck v3, that utilizes machine learning to inform and empower users with respect to privacy policies. PrivacyCheck v3 adds capabilities that are commonly absent from similar PETs. In particular, it adds the ability to (1) find the competitors of an organization with Alexa traffic analysis and compare policies across them, (2) follow privacy policies the user has agreed to and notify the user when policies change, (3) track policies over time and report how often policies change and their trends, (4) automatically find privacy policies in domains, and (5) provide a bird's-eye view of privacy policies the user has agreed to. The new features of PrivacyCheck not only inform users about details of privacy policies, but also empower them to understand privacy policies at a higher level, make informed decisions, and even select competitors with better privacy policies.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**; • **Social and professional topics → Privacy policies**.

KEYWORDS

privacy policy, privacy enhancing technologies, usable privacy, privacycheck

1 INTRODUCTION

Online privacy policies are legal documents that explain how an organization collects, handles, shares, discloses, and uses user data. Privacy policies have grown into the de facto method of communicating such data practices for organizations, and particularly their websites.

Users, however, rarely take the time or effort to read privacy policies [19]. In fact, research shows that only 4.5% of users claim to always read privacy policies [13]. More reliable server-side observation by websites shows that the percentage of users clicking on privacy policies might be as little as 1% [11]. Prior research has demonstrated that the lack of readability in privacy policies is, at least partially, to blame for this lack of interest from users to read them [6]: these policies are lengthy and often require college level education to comprehend [6, 8, 12, 14].

To address the poor readability of privacy policies, an emerging field of research focuses on Privacy Enhancing Technologies (PETs) that summarize and visualize online privacy policies (e.g., Polisis [9], Pribots [10], PolicyLint [2], Privee [34], PrivacyGuide [20], tools from the Usable Privacy project [17], other similar tools and research [7, 35], and our own publicly available PrivacyCheck [26, 28, 30]).

These PETs, however, as we review in Section 2, primarily focus on summarizing the privacy policy itself. They usually lack the capability to provide a higher-level understanding of the landscape of privacy policies or how policies change: How does this policy compare with its competitors? What is the average level of protection privacy policies in a particular sector offer? Where can the user find the same products or services with better data protection? How often does this policy change and has it changed since the user last viewed it or its summary?

In this paper, we introduce the third generation of our PrivacyCheck PET tool. The two previous versions of PrivacyCheck incorporated the use of machine learning models to automatically answer 20 questions about the content of any given privacy policy, ten questions rooted in User Control and another ten in the European General Data Protection Regulation (GDPR). The first two versions were used by about one thousand actual users over the past six years, since the first release in May 2015. We studied the usage patterns of PrivacyCheck v2 [25] to find out PrivacyCheck increased the number of times a user consults privacy policies by

*These authors contributed equally to this research.

80%. Inspired by how real users would like to take advantage of PrivacyCheck, in this paper we add new capabilities to PrivacyCheck v3 to empower users with higher-level understanding of privacy policies. We make the following contributions:

(1) We present the first PET tool to find the competitors of an organization with Alexa traffic analysis and compare policies across them.
(2) In PrivacyCheck v3, we provide the capability to follow privacy policies and notify the user when policies change.
(3) PrivacyCheck v3 tracks policies over time and reports how often policies change and their trends.
(4) PrivacyCheck v3 automatically finds privacy policies in domains.
(5) Our work is the first to provide a bird's-eye view of privacy policies to which the user has agreed.

These additional capabilities will benefit the users of PETs that summarize privacy policies. For example, consider the capability that notifies the user when policies change. Researchers have shown that the majority (63%) of U.S.-based companies only *passively post* new policies online and continuing to use the website indicates users' implicit agreement [27]. Given that these companies do not actively notify users of privacy policy change, our added capability of checking policies frequently and notifying users of change is hugely beneficial.

We organize the rest of this paper as follows. Section 2 covers closely related work and identifies the gap in similar PET tools that we seek to fill. Section 3 provides a brief summary of the previous generations of PrivacyCheck. Section 4 details our new capabilities added to PrivacyCheck and finally Section 5 concludes the paper.

2 RELATED WORK

The flourishing field of PET development has resulted in research and (sometimes publicly accessible) tools that digest long privacy policies and automatically answer questions about them. In this section, we review the most related PET tools and research, with an emphasis on if and how they provide the higher-level picture of privacy policies.

Privee [34] was the first automatic privacy policy analysis tool to utilize machine learning. Building on the crowd sourcing privacy analysis framework ToS;DR [21], Privee combines crowd sourcing with rule and machine learning classifiers to classify privacy policies that are not already rated in the crowd sourcing repository. Privee, however, does not go beyond this basic analysis.

Polisis, available as a web page[1] and a browser extension, utilizes deep learning to summarize what user data privacy policies collect and share. At its core, Polisis is a neural network classifier trained on privacy policies retrieved from the Google Play store. In addition to providing the summary, Polisis visualizes user data collection/sharing, mapping types of data the policy collects/shares to the collection/sharing reasons outlined therein. Furthermore, Polisis displays user choices, security, data retention, etc. as graphs, making it easier for the user to comprehend what is covered in the *current* privacy policy. Notably, Polisis particularly extracts statements about how the policy claims to handle changes in its content. None of these capabilities, nonetheless, go beyond the analysis of the current policy at hand. Even the "policy change" is limited to information extraction from the current privacy policy. Pribots [10] is from the same authors of Polisis and is a chat bot that answers free form questions about a given privacy policy.

The Usable Privacy Project[2] [17] takes advantage of machine learning and crowd sourcing to semi-automatically annotate privacy policies. This project annotates [22, 23] a corpus of 115 policies with attributes and data practices, the same corpus that Polisis and Pribots use to extract coarse- and fine-grained classes.

PolicyLint [2] is a natural language processing tool that identifies potential contradictions that may arise inside the same privacy policy. PrivacyGuide [20] is a machine learning and natural language processing tool inspired by the GDPR. PolicyLint, PrivacyGuide, and many other recently developed tools [3, 4] are solely focused on automatic extraction of information from *one* privacy policy.

Researchers have also investigated the consistency, or lack thereof, between privacy policies of mobile applications and how their actual code treats user data. While research has provided statistics across mobile apps [35, 36] (e.g., 12% of apps handle user location but are silent about it in their privacy policies [35]) these types of statistics are limited to published articles. It is not possible for the user to go to a given privacy policy or app and compare its practices (e.g., location sharing practice) with other apps on demand.

Research that follows privacy policy change over time (e.g., [1, 14]) or examines the landscape of privacy policies with respect to a given criteria (e.g., [5, 18]) has also been popular over the past two decades. While illuminating the high-level picture of how privacy policies change over time or comply with regulation across the board, users of these research projects and their tools are still unable to pick a privacy policy and compare it with others, or draw conclusions about a set of policies that are of importance to them.

At the Center for Identity at the University of Texas at Austin[3] we target many aspects of identity management and privacy [16, 29, 31–33]. We developed PrivacyCheck v1 [30] and v2 [26, 28], as detailed in the next section.

3 BACKGROUND: PRIVACYCHECK

PrivacyCheck is a publicly available browser extension that summarizes privacy policies with machine learning. It automatically answers twenty questions, rooted in the FIPPs (Fair Information Practice Principles) and GDPR (European General Data Protection Regulation). Our previous work covered how we chose these questions and trained LightGBM machine learning models for them (FIPPs questions [30] and GDPR questions [26, 28]). We have also applied PrivacyCheck in a variety of applications: e.g., to study the effect of the GDPR on the landscape of privacy policies [28], to compare privacy policies in the public and private sectors [24], to study privacy policies across industries [27], and to study PET usage patterns [25].

Figure 1 shows the main page of the PrivacyCheck v3 extension. The user navigates to a web page using the Chrome browser and then opens and runs the PrivacyCheck Chrome extension. PrivacyCheck's machine learning models digest the privacy policy to

[1] https://pribot.org/polisis

[2] https://usableprivacy.org

[3] https://identity.utexas.edu

answer ten questions for the (FIPPs-based) User Control and another ten yes/no questions for the GDPR standards. Table 1 lists the User Control and GDPR questions and the scores the privacy policy would receive based on the way it answers each question. The average of the ten scores for each standard is displayed as the score for that standard on the main page (Figure 1): one overall score for the (FIPPs-based) User Control and one for the GDPR. Clicking on each of the scores in Figure 1 takes the user to score breakdowns explaining why the privacy policy received this score. Figures 2 and 3 display the breakdown of the User Control and GDPR scores for a sample website, respectively. See our previous work on the details of the machine learning models (PrivacyCheck v1 [30] and v2 [26, 28]).

PrivacyCheck is currently installed on about 800 Chrome browsers by users around the globe. PrivacyCheck v3 is available online[4] and can also be found by searching for "PrivacyCheck" on the Google Chrome Web Store[5].

PrivacyCheck comprises a front-end that runs on Google Chrome web browser and a back-end that runs on Amazon Web Services (AWS) Lambdas. The front-end takes inputs from the user and sends them to the back-end, and the back-end analyzes the inputs and returns the results to the front-end.

In our prior work [25], we presented a preliminary implementation of a Competitor Analysis Tool (CAT). The main idea was to provide PrivacyCheck users with three other companies in the same market sector as of the policy under evaluation that have received the best scores from PrivacyCheck. However, the way we discovered other organizations in the same market sector proved to provide a questionable fit: (1) There were a total of only 15 market sectors so many organizations with a variety of services would be coarsely grouped in the same market sector and see the *same* best competitors over and over again. For instance, the lack of a sufficient amount of sectors led the old CAT to group website URLs like digitaltruth.com, allybaggett.net, dogonvillage.com, airbnb.com, google.com, and facebook.com into the same group, "Computers". (2) The machine learning classifiers we used to "guess" the market sectors of organizations had to compromise the accuracy of classification for efficiency and availability, achieving an accuracy of only 55%. In this paper, we revamp the CAT component of PrivacyCheck to provide a finer grade and more accurate set of competitors based on Alexa's traffic analysis (Section 4). Because we use Alexa Competitive Analysis offered by Amazon[6], the accuracy of competitor analysis in PrivacyCheck v3 is virtually 100%—there is no need for a classification model. Furthermore, the competitive analysis is very fine grained, as there is no limit to a number of classes for classification.

4 NEW CAPABILITIES OF PRIVACYCHECK V3

In this section, we detail each of the new capabilities of PrivacyCheck as follows.

(1) PrivacyCheck v3 finds privacy policies and scores them, without having the user to manually navigate to the privacy policy page of a website.
(2) As many organizations and companies have multiple privacy policies, each governing one of their services/products, PrivacyCheck v3 provides the option to aggregate all privacy policies across an organization.
(3) PrivacyCheck v3 tracks privacy policy score data over time, thus it can (1) display the historical scores of a privacy policy over time on a graph for users and (2) notify users when there is a change in a privacy policy.
(4) PrivacyCheck v3 shows the score distribution of the privacy policies to which the user has agreed, in order to reflect the user's overall privacy risk level.
(5) Finally, PrivacyCheck v3 improves the competitor analysis to provide more relevant competitors with very fine-grained categories.

4.1 Automatically Finding Privacy Policies

PrivacyCheck previously needed the user to navigate to a privacy policy page and re-run the tool manually in order to view a policy's score. We eliminated this need by automatically finding privacy policies of any given domain, including those on the domain as well as privacy policies on other domains if this web page links to them. For example, finding the privacy policy of Youtube.com will also return Youtube's privacy policy on Google's domain, which is the correct privacy policy.

Given the URL of the user's browser tab, PrivacyCheck v3 scrapes the current page for new policies and then scores them. Using Beautiful Soup, a Python web scraping library, PrivacyCheck scrapes for all the links on the main page of the domain. To determine if a link is a privacy policy, we take advantage of prior research [15, 18, 35] indicating that the URLs of links to privacy policies commonly have a certain set of keywords, including "privacy", "legal", "conditions", "policy", "policies", "terms", and "info".

Figure 4 depicts the button to find and score new policies. These results are divided into two sections; policies hosted on the same domain and policies hosted on another domain. For example, Youtube contains links to privacy policies that are located on both Google's website and its own website. These policies would be found and scored, but shown separately (Figure 5).

4.2 Aggregating Scores Across All Privacy Policies of a Domain

Additionally, some websites have many different privacy policies depending on the size and scale of the organization. If a user wants to have a better understanding of the privacy strength of an organization, they should not have to manually navigate and run the extension on various different web pages, as was the case in the previous versions of PrivacyCheck.

In addition to the button to find and score new policies, Figure 4 also shows the aggregate panel, which contains a table of all of the privacy policy scores for the user's current website already examined by *any* PrivacyCheck user. When the user wishes to view existing scores or find new ones, they simply navigate to this panel and all of the scores are fetched and displayed. For example, if the user is currently browsing apple.com, the new panel contains a table with related policies such as apple.com/legal.

[4]https://tinyurl.com/ydf7h7dr
[5]https://chrome.google.com/webstore
[6]https://www.alexa.com

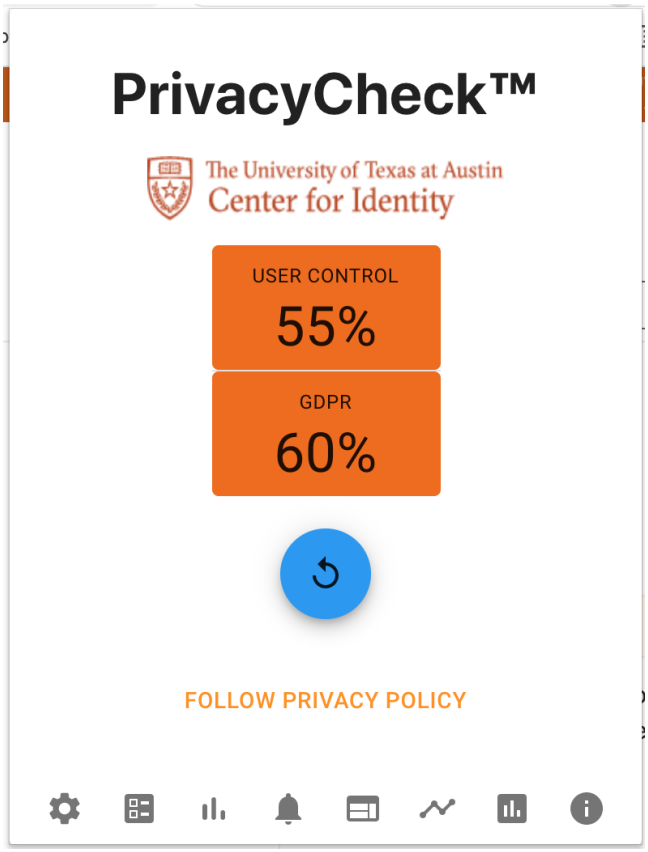

Figure 1: PrivacyCheck v3.

Table 1: PrivacyCheck v2 questions and scoring methods, kept in v3.

	User Control	Scores: 100% (Green)	50% (Yellow)	0% (Red)
1	How well does this website protect your email address?	Not asked for	Used for the intended service	Shared w/ third parties
2	How well does this website protect your credit card information and address?	Not asked for	Used for the intended service	Shared w/ third parties
3	How well does this website handle your social security number?	Not asked for	Used for the intended service	Shared w/ third parties
4	Does this website use or share your PII for marketing purposes?	PII not used for marketing	PII used for marketing	PII shared for marketing
5	Does this website track or share your location?	Not tracked	Used for the intended service	Shared w/ third parties
6	Does this website collect PII from children under 13?	Not collected	Not mentioned	Collected
7	Does this website share your information with law enforcement?	PII not recorded	Legal docs required	Legal docs not required
8	Does this website notify or allow you to opt-out after changing their privacy policy?	Posted w/ opt out option	Posted w/o opt out option	Not posted
9	Does this website allow you to edit or delete your information from its records?	Edit/delete	Edit only	No edit/delete
10	Does this website collect or share aggregated data related to your identity or behavior?	Not aggregated	Aggregated w/o PII	Aggregated w/ PII
	GDPR	Scores: 100% (Green)		0% (Red)
1	Does this website share the user's information with other websites only upon user consent?	Yes		No/Unanswered
2	Does this website disclose where the company is based/user's PII will be processed & transferred?	Yes		No/Unanswered
3	Does this website support the right to be forgotten?	Yes		No/Unanswered
4	If they retain PII for legal purposes after the user's request to be forgotten, will they inform the user?	Yes		No/Unanswered
5	Does this website allow the user the ability to reject usage of user's PII?	Yes		No/Unanswered
6	Does this website restrict the use of PII of children under the age of 16?	Yes		No/Unanswered
7	Does this website advise the user that their data is encrypted even while at rest?	Yes		No/Unanswered
8	Does this website ask for the user's informed consent to perform data processing?	Yes		No/Unanswered
9	Does this website implement all of the principles of data protection by design and by default?	Yes		No/Unanswered
10	Does this website notify the user of security breaches without undue delay?	Yes		No/Unanswered

In order to implement this feature, we took advantage of the fact that PrivacyCheck maintains a database of privacy policy summaries on the server side. PrivacyCheck hosts this DynamoDB database on AWS and utilizes it 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, GDPR and User Control scores, and the date that the

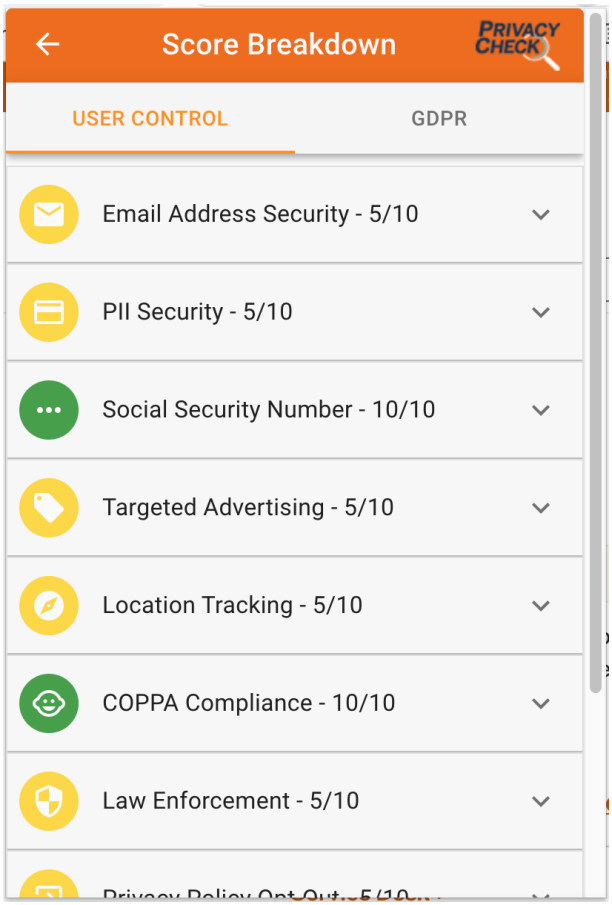

Figure 2: The breakdown of the User Control score for a sample website.

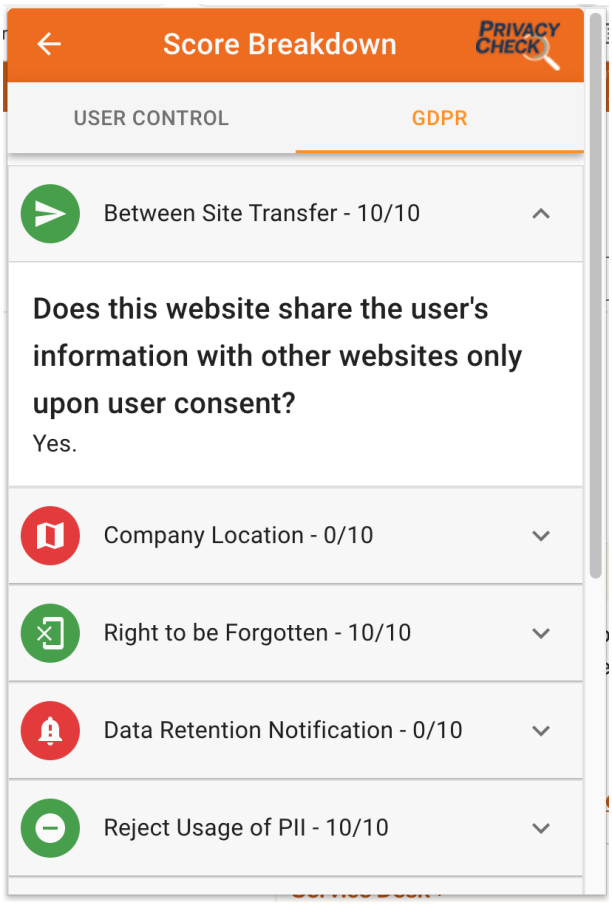

Figure 3: The breakdown of the GDPR score for a sample website.

entry was made. Every time PrivacyCheck is run on a new privacy policy, an entry is added to the database. If PrivacyCheck is run on a policy that already exists in its database, it merely updates the corresponding entry. Searching through this database (which does not contain any personal information from PrivacyCheck users) enables PrivacyCheck v3 to readily identify all privacy policies and scores for the user's current website domain already examined by *any* PrivacyCheck user.

4.3 Tracking Privacy Policies over Time

By tracking privacy policies over time, we pursue two goals: (1) to notify users of a privacy policy change, and (2) to visualize how frequently a privacy policy changes and if it is trending towards improving or worsening. In order to provide these capabilities, we updated the front-end of PrivacyCheck as well as its back-end lambda and database.

First, we add to the front-end a feature that the user would utilize to indicate they have consented to a privacy policy. We should not assume the user is consenting to the privacy policy of every single website they visit, as they might review the privacy policy and decide not to provide consent. On the other hand, not every website actively displays a button that reads "I agree": for many websites, the continued use indicates implicit consent. Therefore, to cover all the various ways of providing consent common on the Internet, we added a feature to the front-end that the user would use to indicate they have agreed to a privacy policy. When a user runs PrivacyCheck on a privacy policy, PrivacyCheck gives them the option to add the policy to the "following list", i.e., the list of privacy policies they have agreed to and are interested in following. Figure 6 shows the following list tab, which displays the latest scores and dates for a policy, along with the option to remove the policy from the list. Alternatively, clicking the bell icon on the bottom of the main page of PrivacyCheck takes the user to the following list.

Second, tracking the policy changes over time requires modifications in the database structure and the back-end lambda code of PrivacyCheck too. We keep all the privacy policies examined with PrivacyCheck in a central DynamoDB NoSQL database on AWS, excluding the information about the users who examined them to protect the privacy of PrivacyCheck users. Whenever a change is detected, we update the data kept about each privacy policy dynamically with new scores and dates.

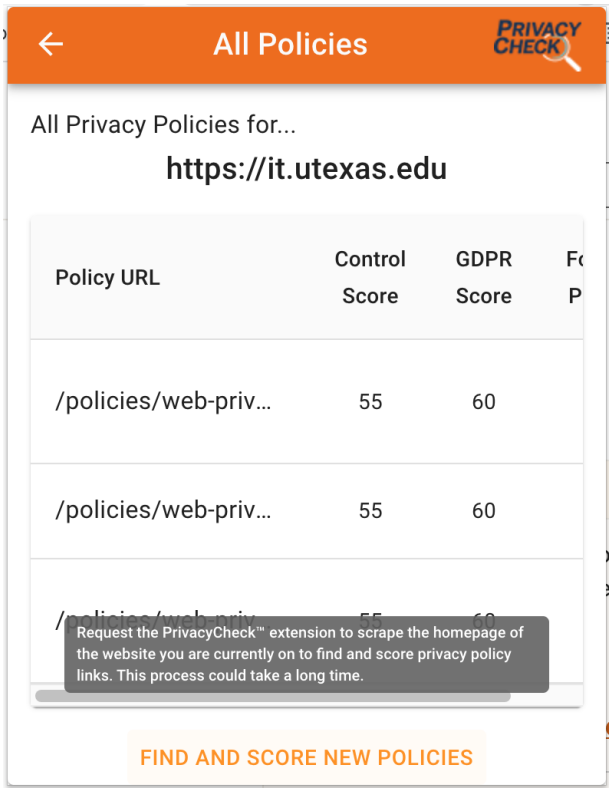

Figure 4: Finding and scoring new policies on the domain of a sample website.

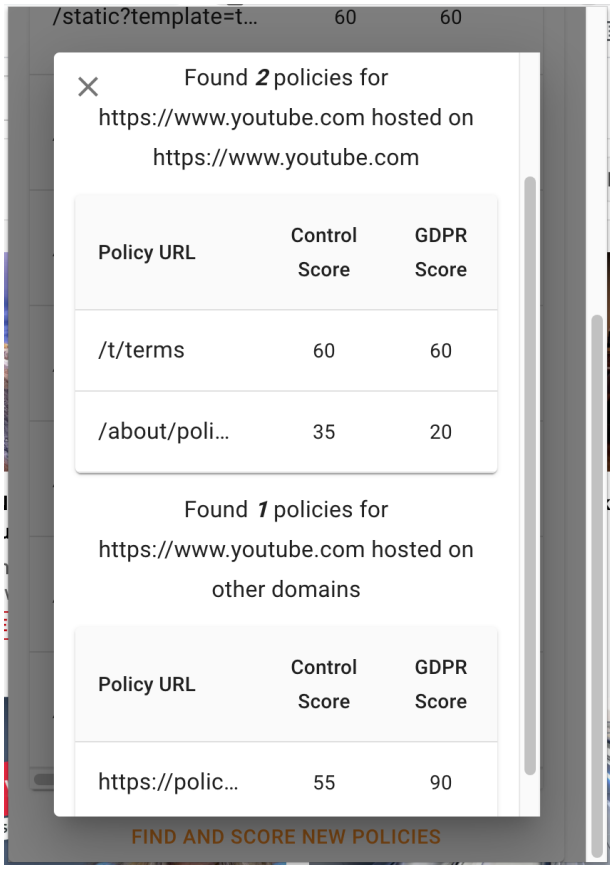

Figure 5: Policies found for Youtube on its own domain and other domains.

To protect the privacy of PrivacyCheck users, the list of policies they follow is stored in the local storage only. The local PrivacyCheck extension has this list along with the most recently seen score. Every time the extension is launched, a query is sent to the database back-end for each followed privacy policy, asking for the most up to date scores. If the extension realizes there is a more up to date score than what it currently has, it displays a notification icon to alert the user that the score for that followed policy has changed.

Furthermore, the user can run PrivacyCheck on any of the policies in the following list, regardless of their current browser page, and navigate to the score breakdowns and historical data for the policy. This removes the need to traverse to a specific privacy policy website to run PrivacyCheck. Figure 7 is the line graph that shows historical scores for the policy under examination. Hovering over each data point shows more details including the date of that score.

4.4 Privacy Policy Score Distribution

The score distributions panel displays a new and interactive chart interface that graphically displays the privacy score of all the privacy policies that the user is following. This capability allows a user to have a more concrete understanding of privacy policies they follow. This understanding can help users make changes to their behavior online in order to protect their data. A snippet of the privacy policy URL is shown on the side to identify each graph, and upon hovering on one of the bars, the user is able to view the full privacy policy URL. This page also breaks down the scores and separately displays both User Control and GDPR scores so that the user is able to distinguish between the two. Figure 8 displays the score distribution tab.

Recall that PrivacyCheck's machine learning models (inherited from v2) digest the privacy policy and assign two scores to it, one for the (FIPPs-based) User Control and one for the GDPR standards. Each of these scores is the average of the ten scores this particular privacy policy received because of the way it collects and handles user data according to Table 1. See our previous work on the details of the machine learning models (v1 [30] and v2 [26, 28]).

4.5 Improved and Fine-Grained Competitor Analysis

As mentioned in Section 3, the previous competitor analysis tool of PrivacyCheck v2 was insufficient for several reasons such as being too generic and inaccurate. We replaced the low accuracy machine learning models of the CAT tool with the Alexa Competitive

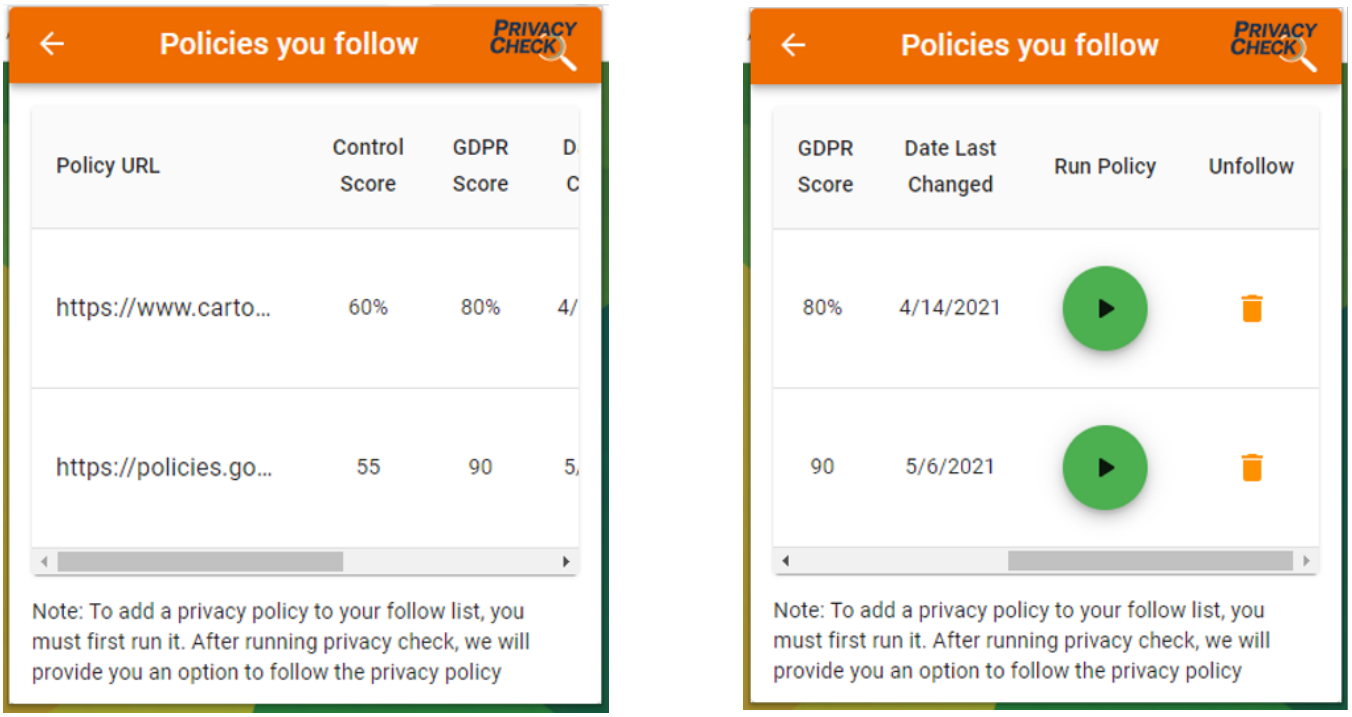

Figure 6: The following panel.

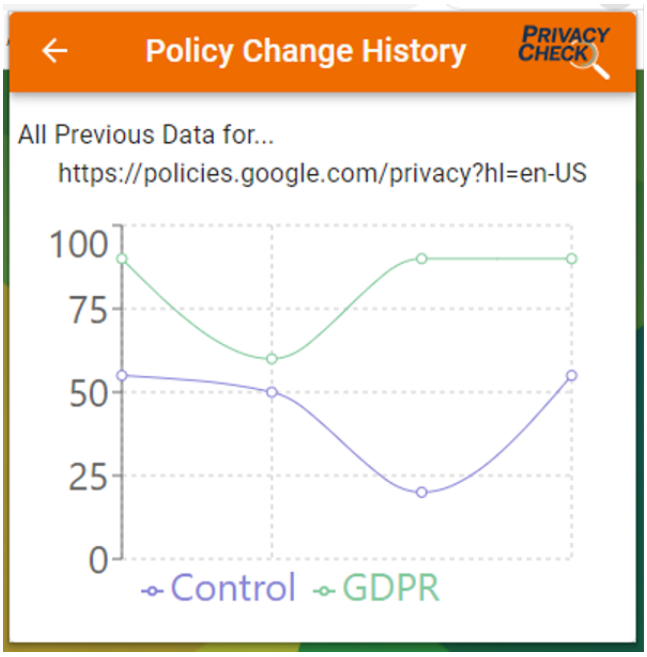

Figure 7: The historical data of User Control and GDPR scores for a sample policy.

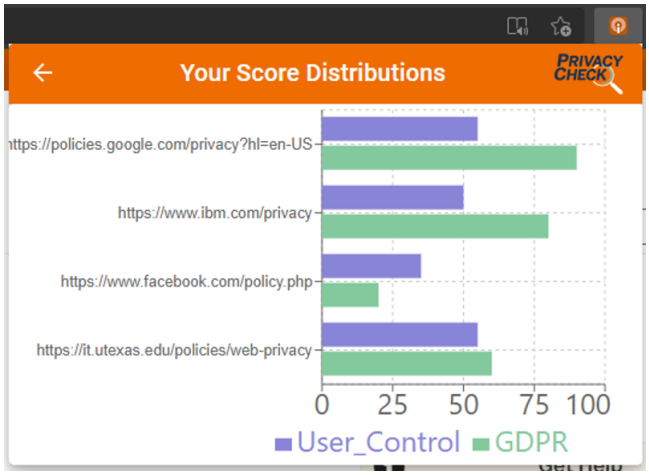

Figure 8: The User Control and GDPR score distribution of the policies that the user is following.

Analysis offered by Amazon[7]. Alexa is a web traffic analysis tool that returns the top five competitors when given a domain URL, based on the frequency of websites being retrieved together, shared keywords, and search results. PrivacyCheck's back-end calls Alexa with the domain of the privacy policy under investigation, and uses Alexa's five competitors results to fetch the privacy policies of competitors and calculate their scores. Then, the back-end sends the results to the front-end to display the competitors and their scores as a list, with separate tabs for GDPR and User Control scores.

[7]https://www.alexa.com

In PrivacyCheck v2 [25], the main idea of CAT was to provide PrivacyCheck users with some other companies in the same market sector as of the policy under evaluation that received the best scores from PrivacyCheck. To find competitors, we used machine learning: a classification model with 15 market sectors. The machine learning classifier in v2 used to *guess* the market sectors. It also had to compromise the accuracy of classification for efficiency and availability, achieving an accuracy of only 55%. In PrivacyCheck v3, we use Alexa Competitive Analysis in lieu of the classifier: the accuracy of competitor analysis in PrivacyCheck v3 is virtually 100%—there is no need for a classification model to guess competitors. Very fine-grained competitors are readily provided by Alexa through shared keyword and traffic analysis and no limit exists for the number of classes.

Figure 9 depicts the competitors for the web site of the University of Texas at Austin, along with their color-coded User Control and GDPR scores. Green, yellow, and red show low, medium, and high compliance respectively. The links to competitors are also included for easy access.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we covered the newest generation of PrivacyCheck—a research project and browser extension meant to make comprehending privacy policies easier. We added four capabilities to PrivacyCheck: automatically finding privacy policies, following them and getting notified about their changes, tracking them over time, and the high-level landscape of all the policies to which the user has agreed. In addition, we improved PrivacyCheck's competitor analysis tool to support very fine-grained discovery of competitors through Amazon Alexa. Using PrivacyCheck v3, users can realize how privacy policies to which they have agreed change over time, as well as become aware of the high-level picture of these policies. Finally, PrivacyCheck enables and empowers users to make informed decisions and even switch to product and service providers with better privacy policies.

For future work, we envision answering multiple sets of questions with PrivacyCheck:

5.1 Improving PrivacyCheck Questions and Their Machine Learning Models

How can PET tools improve their accuracy by training on millions of privacy policies available online? We will add new questions to PrivacyCheck and train its machine learning models on emerging very large datasets of privacy policies [15, 18] to answer these new questions.

5.2 Enhancing the Usability of PET Tools

How can we measure the interest of the public in PET tools like PrivacyCheck? How can we improve the performance of the current capabilities of PrivacyCheck and incorporate the feedback we receive from the actual users of PrivacyCheck?

5.3 Measuring PET Traffic

What is the traffic of PrivacyCheck usage like? How do users interact with PET tools like PrivacyCheck? Do users have a tendency to run PET tools on a particular type of policies and/or in specific sectors? What are fine-grained categories of the privacy policies that are of the most interest to users? Is there a statistically significant relationship between the length, readability, type of data, or market sector of privacy policies with the frequency of users running PET tools on them?

5.4 Measuring PET Usage Patterns

How do users interact with the summary that the PET tool provides? How much time do they invest in reading the summary and how does that time fair against actually reading the entire policy?

ACKNOWLEDGMENTS

This work was in part funded by the Center for Identity's Strategic Partners. The complete list of Partners can be found at https://identity.utexas.edu/strategic-partners.

REFERENCES

[1] 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).

[2] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. Policylint: investigating internal privacy policy contradictions on Google play. In *28th USENIX Security Symposium*. 585–602.

[3] Vanessa Bracamonte, Seira Hidano, Welderufael B Tesfay, and Shinsaku Kiyomoto. 2020. Evaluating the Effect of Justification and Confidence Information on User Perception of a Privacy Policy Summarization Tool.. In *ICISSP*. 142–151.

[4] Duc Bui, Kang G Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated Extraction and Presentation of Data Practices in Privacy Policies. *Proceedings on Privacy Enhancing Technologies* 2021, 2 (2021), 88–110.

[5] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy... Now Take Some Cookies-Measuring the GDPR's Impact on Web Privacy. *Informatik Spektrum* 42, 5 (2019), 345âĂŞ346.

[6] Tatiana Ermakova, Annika Baumann, Benjamin Fabian, and Hanna Krasnova. 2014. Privacy Policies and Users' Trust: Does Readability Matter?. In *20th Americas Conference on Information Systems (AMCIS)*.

[7] Kassem Fawaz, Thomas Linden, and Hamza Harkous. 2019. The Applications of Machine Learning in Privacy Notice and Choice. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*. IEEE, 118–124.

[8] Mark A Graber, Donna M D Alessandro, and Jill Johnson-West. 2002. Reading level of privacy policies on internet health web sites. *Journal of Family Practice* 51, 7 (2002), 642–642.

[9] 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 *27th USENIX Security Symposium*. 531–548.

[10] Hamza Harkous, Kassem Fawaz, Kang G Shin, and Karl Aberer. 2016. Pribots: Conversational privacy with chatbots. In *Twelfth Symposium on Usable Privacy and Security (SOUPS) 2016*.

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

[12] Aleecia M McDonald and Lorrie Faith Cranor. 2008. the Cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society* 4 (2008), 543.

[13] George R Milne and Mary J Culnan. 2004. Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of Interactive Marketing* 18, 3 (2004), 15–29.

[14] George R Milne, Mary J Culnan, and Henry Greene. 2006. A longitudinal assessment of online privacy notice readability. *Journal of Public Policy & Marketing* 25, 2 (2006), 238–249.

[15] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2021. A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*. 143–148.

[16] Rima Rana, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2019. An Assessment of Blockchain Identity Solutions: Minimizing Risk and Liability of Authentication. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 26–33.

[17] Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonalda, Joel R Reidenbergb, Noah A Smith, Fei Liu, N Cameron Russellb, Florian Schaub, et al. 2013. *The usable privacy policy project*. Technical Report. Technical Report, CMU-ISR-13-119, Carnegie Mellon University.

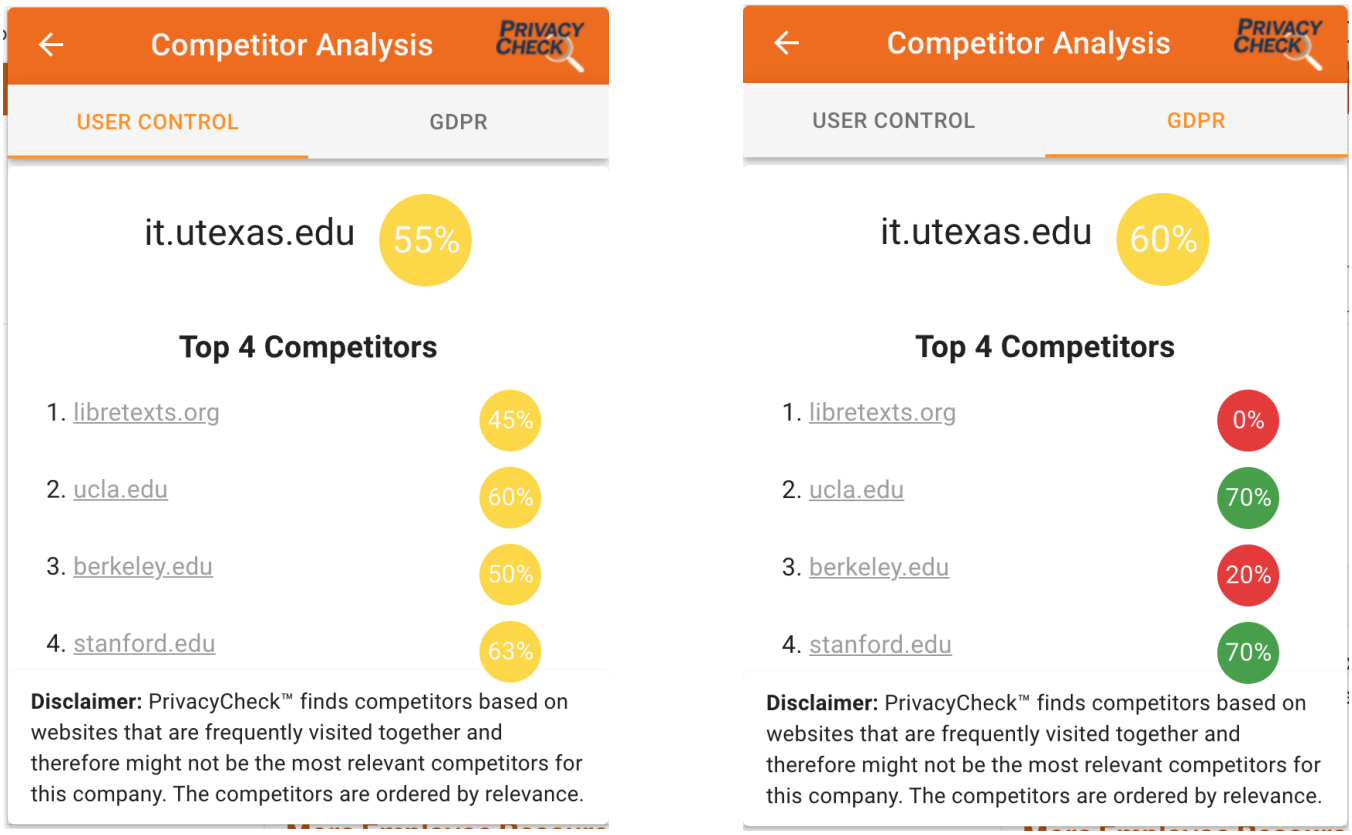

Figure 9: The competitor analysis for a sample web site—User Control scores (left) and GDPR scores (right).

[18] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2020. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. *arXiv preprint arXiv:2004.11131* (2020).

[19] Nili Steinfeld. 2016. âĂIJI agree to the terms and conditionsâĂİ:(How) do users read privacy policies online? An eye-tracking experiment. *Computers in human behavior* 55 (2016), 992–1000.

[20] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. 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, 15–21.

[21] ToS;DR. 2012. Terms of Service; Didn't Read. https://tosdr.org

[22] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Annual Meeting of the Association for Computational Linguistics*. 1330–13340.

[23] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A Smith, and Frederick Liu. 2016. Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?. In *Proceedings of the 25th International Conference on World Wide Web*. 133–143.

[24] Razieh Zaeem and K. Barber. 2021. Comparing Privacy Policies of Government Agencies and Companies: A Study using Machine-learning-based Privacy Policy Analysis Tools. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART,*. INSTICC, SciTePress, 29–40. https://doi.org/10.5220/0010180700290040

[25] Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, and K Suzanne Barber. 2020. PrivacyCheck's Machine Learning to Digest PrivacyPolicies: Competitor Analysis and Usage Patterns. In *The 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'20)*. To Appear.

[26] Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K. Suzanne Barber. 2020. PrivacyCheck v2: A Tool that Recaps Privacy Policies for You. In *29th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM. To appear.

[27] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2017. A study of web privacy policies across industries. *Journal of Information Privacy and Security* 13, 4 (2017), 169–185.

[28] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2020. The effect of the GDPR on privacy policies: recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)* 12, 1 (2020), 1–20.

[29] Razieh Nokhbeh Zaeem, Suratna Budalakoti, K Suzanne Barber, Muhibur Rasheed, and Chandrajit Bajaj. 2016. 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, 1–8.

[30] Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. 2018. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Transactions on Internet Technology (TOIT)* 18, 4 (2018), 53.

[31] Razieh Nokhbeh Zaeem, Monisha Manoharan, and K Suzanne Barber. 2016. Risk kit: Highlighting vulnerable identity assets for specific age groups. In *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 32–38.

[32] Razieh Nokhbeh Zaeem, Monisha Manoharan, Yongpeng Yang, and K Suzanne Barber. 2017. Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Computers & Security* 65 (2017), 50–63.

[33] Jim Zaiss, Razieh Nokhbeh Zaeem, and K Suzanne Barber. 2019. Identity Threat Assessment and Prediction. *Journal of Consumer Affairs* 53, 1 (2019), 58–70.

[34] Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *23rd USENIX Security Symposium*. USENIX Association, San Diego, CA, 1–16.

[35] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies* 2019, 3 (2019), 66–86.

[36] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven Bellovin, and Joel Reidenberg. 2016. Automated analysis of privacy requirements for mobile apps. In *2016 AAAI Fall Symposium Series*.

WWW.IDENTITY.UTEXAS.EDU