Reproducibility of Web Privacy Measurements

Karel Kubicek
karel.kubicek@inf.ethz.ch
ETH Zurich
Zurich, Switzerland

Ahmed Bouhoula
ahmed.bouhoula@inf.ethz.ch
ETH Zurich
Zurich, Switzerland

Patrice Kast
patrice@kasts.ch
ETH Zurich
Zurich, Switzerland

ABSTRACT

The collection of private data engages in a continuous cat-and-mouse game between the advertising industry and privacy regulators. Existing web privacy studies expose a high prevalence of privacy-intrusive techniques, yet most are confined to one-time measurements. Challenges such as a lack of documentation, limited code availability, and software aging create obstacles to reproducibility, hindering researchers from assessing web privacy trends over time. This knowledge gap obstructs the impact assessment of privacy-enhancing technologies and poses a hurdle to effective policymaking.

Our project introduces the Privacy Observatory framework to tackle these challenges. By collecting, systematizing, and reproducing privacy measurement studies, it consolidates results, analyzes correlations in advertising and tracking methods, and assesses the evolution of privacy practices over time. Utilizing virtualization and orchestration for unified large-scale analysis, the framework addresses reproducibility issues. Our evaluation on five studies demonstrates the framework's effectiveness and allows us examining reproducibility in practice. Our results, providing a comprehensive understanding of web privacy trends, are designed to empower stakeholders to make informed decisions shaping the future of online privacy.

1 INTRODUCTION

Modern websites heavily rely on collecting and processing private data for advertising. These practices are regulated by laws such as the General Data Protection Regulation (*GDPR*) or ePrivacy Directive. Despite regulations, the advertising and tracking industry, valued at $200 billion [17], has an advantage in the imbalanced cat-and-mouse play between firms and data protection authorities (*DPAs*).

Empirical researchers conduct web privacy measurements to observe privacy-intrusive practices, such as cookie tracking [28, 32] or browser fingerprinting [15, 22]. Similarly, some studies measure compliance with regulations [4, 5, 21, 24]. While these studies warn of a high spread of privacy-intrusive techniques and their non-compliance, the majority of them are only one-time measurements. Challenges, such as a lack of documentation and code [11, 35] or rapid software aging [19, Sec. 5.1], implies high complexity for reproducibility. This, together with the lack of incentives for reproduction studies and the incompatibility among various measurement methods, hinders researchers from assessing the web privacy time-trends and from understanding the mutual state of tracking techniques. In the current situation, the authors of privacy-enhancing technologies and policymakers lack feedback regarding the impact of their work, hindering future decision-making.

Our work. Our project aims to collect, systematize, and most importantly reproduce privacy measurement studies, consolidate results on a unified sample, and analyze correlations in advertising and tracking methods. Additionally, it seeks to measure the evolution of privacy practices over time, evaluating the impact of new technologies and laws on end-user privacy. In other words, the project aims to shed light on the overarching question: *How are we doing in protecting privacy?*

To efficiently analyze the publications in scale, we developer Privacy Observatory framework that uses virtualization of individual study implementations and provides orchestration capabilities for simple deployment using shared crawling lists and simple collection of the results.

We evaluated the capabilities of Privacy Observatory and the process of reimplementation on five exemplary studies. We were able to reproduce four of these works, and the time required to do so ranged from 16 to 41 hours. We show that source code artifacts from published studies are not designed to guarantee long-term functionality or were not thoroughly tested. Especially the post-processing step, where data from individual crawls are aggregated according to measurement scenarios, procedures are significantly less documented, let alone automated, or is completely missing. We found that while some of these studies followed the principles by Demir et al. [11], they still face severe troubles with reproducibility, but if they would use containers, six criteria would be satisfied by default and further six can be achieved by using the features our Privacy Observatory.

Contributions. In addition to the creation of Privacy Observatory framework, our work will contribute in the following ways.

- It will observe whether the web measurement field faces a reproducibility crisis similarly to other fields [2].
- By comparing old and new results, it will allow us to reason about the evolution in web privacy.
- It will identify and limit biases caused by website sampling, as Ruth et al. [29] reported.
- It will allow us to report on potential correlations among privacy-intrusive practices or privacy violations.

2 RELATED WORK

In this section, we first illustrate the web measurement studies and their (lack of) focus on time-trend analysis, then we present reproducibility studies as an alternative to long-term studies, and finish with systematization studies in the field of applied privacy.

Scope. The field of web privacy and privacy compliance measurements is currently experiencing dynamic growth. Every year, approximately 50-100 of new privacy (compliance) measurements are published. These publications serve as the foundation for our

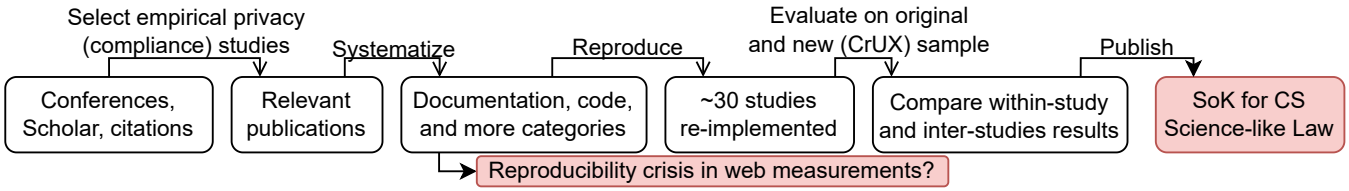

Figure 1: Overview of steps of our study and results.

data collection. Below, we present the diverse results of prior research on example of cookie notices, a highly active aspect of web compliance measurements.

The analysis of cookie consent compliance requires detection of cookie notices, which has been undertaken by various publications. Degeling et al. [10] identified cookie notices on 62% of websites in 2018, Sachez-Rola et al. [30] observed them on 57% in 2019, Kampanos et al. [18] identified them on 45% in 2021, and Bouhoula et al. [5] observed cookie notices on 57% of analyzed websites. While an apparent downward trend may be inferred from these figures, such an interpretation is counterintuitive in the context of gradual website adoption of legal requirements and inconsistent with user experiences. The discrepancy across these publications arises from variations in detection methods and sampled websites, biasing the results and rendering single-time measurements unsuitable for longitudinal trend analysis. Consequently, such studies, predominant in web measurement literature, fail to provide valuable insights to policymakers and developers of privacy-enhancing technologies.

Notably, several web measurements offer a broader temporal perspective. Trevisan et al. [34] conducted a four-year study on cookie compliance, albeit limited to a dataset of only 241 websites. Degeling et al. [10] investigated the presence of cookie notices and privacy policies for nearly a year around the enactment of GDPR. Furthermore, studies utilizing archived datasets, like Amos et al. [1], analyzed privacy policies from WebArchive spanning more than two decades. However, while such archives are suitable for examining declared privacy behavior (e.g., the text of privacy policy), they are inadequate for observing actual private data collection practices.

Reproducibility. An alternative method to observe time-trends, together with avenues for comparison of various methods, is reproducing published studies. Olszewski et al. [26] investigated the reproducibility of 750 machine-learning publications. Their work propose five recommendations to artifact evaluation committees, such as requiring self-contained environments such as Docker images. Demir et al. [11] established 18 criteria contributing to the reproducibility of web measurement studies, forming a foundational basis for our reproducibility study, wherein we intend to assess all publications against these criteria. However, we recognize the limitations of these theoretically-established criteria, as discussed in our prior study [19, Chap. 5]. A subsequent publication by Demir et al. [12] examines various experiment configurations but is confined to their own web crawler, without involving the reproduction of published experiments.

Systematization. Several Systematization of Knowledge (*SoK*) publications have systematized knowledge about web and mobile privacy. Mayer et al. [25] summarized third-party web tracking, yet this twelve years old publication in this dynamic field is outdated. Reitinger et al. [27] proposed a meta-study of web compliance publications, without an intention to reproduce the studies they review. Lastly, Birrell et al. [3] conducted an SoK of computer-science (*CS*) publications focusing on privacy laws, categorizing publications into 33 legal aspects across privacy laws (e.g., right to delete or data minimization principle). They conclude with recommendations for CS researchers, namely Recommendations 3 advocate for increased attention to long-term studies in this domain.

3 STUDY DESIGN

The first essential step of our study is to finding publications measure privacy (compliance) of websites. Then the selected publications are systematized according to specified categories. We propose automation for both of these steps to prevent mistakes stemming from the tedious work, but we also plan to manually evaluate sample for assessing the trustworthiness of the results.

3.1 Publication selection

We plan to adhere to the typical publication selection criteria standardized in multiple SoK publications. Our discovery process will utilize the following sources.

(1) All publications from top security and privacy or internet measurement conferences. The applicable period is still to be determined. These conferences include:
   (a) IEEE Symposium on Security and Privacy ('Oakland')
   (b) USENIX Security Symposium
   (c) Network and Distributed System Security Symposium (NDSS)
   (d) ACM Conference on Computer and Communications Security (CCS)
   (e) Privacy Enhancing Technologies Symposium (PETS)
   (f) Symposium on Usable Privacy and Security (SOUPS)
   (g) ACM Conference on Human Factors in Computing Systems (CHI)
   (h) ACM The Web Conference (WWW)
   (i) ACM International Conference on Web Search and Data Mining (WSDM)
   (j) Internet Measurement Conference (IMC)
(2) Semantic Scholar search using a query yet to be determined.
(3) After a filtering step described below, we will consider one level backward and forward citations – papers cited in the selected publications or papers citing them.

We plan to employ an automated approach to this step. Currently, we are developing a proof-of-concept tool to scrape publications from conference pages and use the Semantic Scholar API to search for a selected query or for forward and backward citations. The tool also transforms the publications into a machine-readable text either by by collecting HTML-version of the article when available[1] or using OCR-based library scanning the PDF. Although we are not aware of such automation being used in similar publications, we find it necessary based on the experience with the erroneous process gained during the pilot study performed in [20, Sec. 6.2.1] and mistakes identified in the sampling process in other SoK publications. However, we also plan to manually inspect a portion of the dataset to evaluate the accuracy of this automation.

We expect this process to generate a dataset with up to tens of thousands of publications. For that reason, we plan to automate the filtering process. We will classify titles and abstracts using a large language model as either privacy (compliance) web measurement publications or others that will be discarded. The same filtering will be applied before step 3 of the publication discovery to limit the number of considered publications. We also plan to evaluate the accuracy of the filtering step.

3.2 Systematization

The selected publications are then the input to the systematization step. We plan to first take a small sample based on which we will create systematization criteria. Given the pilot study and other SoKs, we are already aware that we will extract at least the following aspects.

- **Presence of the artifact:** If the artifact is present, which badges have been assigned to the artifact (badges typically include: artifact available, functional, and reproduced), and whether the artifact was awarded any prizes.
- **Presence of the source code:** A publication can release an artifact without source code or it can release code independently of the artifact. The code can also be available only upon request.
- **Reproducibility criteria by Demir et al. [11]:** The text of the paper and the released code would be evaluated according to Demir's 18 criteria that influence the study reproducibility.
- **Privacy or compliance:** Whether the publication have studied compliance with privacy regulations or the privacy loss without connection to any law. When a legal framework is provided, which laws are referenced?
- **Measured privacy aspect:** The aspects are to be finalized according to the sample evaluation, but they will include categories such as cookies, emails, fingerprinting, etc.
- **Sample of websites:** How have the authors sampled websites and whether they have published the original sample.

We also plan to discuss the proposed systematization aspects with influential authors in the area.

Once the systematization aspects are finalized, we plan to instruct a large language model to process the papers and extract the systematization for us. However, if this step proves to be erroneous, we will resolve the task by manually reading the publications. Nevertheless, a sample will be evaluated to analyze the performance of the automation methods, as well as the inter-annotator agreement.

4 REPRODUCTION AND MEASUREMENTS

We will evaluate the reproducibility issues of a selected impactful 30 publications with available code. We developed the Privacy Observatory platform to simplify the maintenance of long-term studies. This platform streamlines the process of regularly repeated crawls and their respective post-processing in order to verify long-term trends over the years with a stable configuration of the study environment. For a detailed reference about the platform, please refer to Patrice Kast's Master's thesis [19].

The central idea of our approach focuses on the use of container images, in particular Docker Images.[2] Source code that is not embedded within a container leaves space for several factors negatively influencing reproducibility, such as variety in deployment configuration or versions of used programs and libraries. In contrast, Docker containers fixes the source code and execution environment at the point of the image build. This is essential for web crawling which has external dependencies, e.g., the existence of old compatible packages in software repositories. In addition, containers simplify the effort to keep a clean browsing environment for each crawl, further reducing potential sources of bias like temporary files or collected cookies which have a negative impact on the reproducibility efforts.

We document the process of embedding studies in a container, and evaluate its impact Demir's reproducibility criteria. These containers are then executed by the Privacy Observatory platform, which is responsible for their scheduling, deployment, providing the input (i.e., crawling list), and collects the outputted measurements.

Using the implementation of the reproduced studies in our framework, we will measure the studied aspect on two samples of websites: a sample from the Chrome User Experience Report (*CrUX*), which better represents website popularity than alternative lists [29] and the original sample from the publication or a comparable sample when applicable. These measurements will allow us to reason about two types of results: the comparison of observations within the same study and the comparison of observations across multiple studies.

Within-study comparison. We will compare, first, the published results with the new results on the original websites sample, second, the new results on the original sample with the new results on the CrUX sample, and third the old results with the new CrUX results. These three comparisons will allow us to discover the trend of the observation as well as the bias introduced by the publications due to the sampling strategy. Such biases might reveal systematic issues in the literature, as predicted by Ruth et al. [29].

Across-studies comparison. Using the CrUX sample, we will inspect the correlations among observations. We will be able to reason

[1] Services such as arXiv, IEEExplore, Springer, or ACM DL offer HTML-based version of the article either through an API or on a website that can be scraped.

[2] Container images are similar to virtual images – they are self-contained and able to run at any host machine. Using technologies such as Docker, these containers are light weight, both in storage requirements and in computational overhead.

Table 1: Summary of the time needed to reproduce the measurements, whether we reproduced the work (column S), the reused components, and the number of satisfied Demir's et al. [11] criteria before (#CB) and after (#CA) our reproduction.

Pub.	Time	S	Reused components	#CB	#CA
[4]	16h	✓	Procedure, framework, and post-processing	13	14
[14]	51h	✗	–	9	15
[6]	41h	✓	Procedure, framework, and post-processing	12	15
[23]	17h	✓	Procedure	8	16
[31]	36h	✓	Procedure, framework, and post-processing		

about important scientific questions such as whether cookie tracking and browser fingerprinting are used mutually or together, with the hypothesis being the latter.

5 PILOT RESULTS

We reproduce the following five influential privacy measurement studies, evaluating both the suitability of our Privacy Observatory platform and the reproducibility of these studies.

Bollinger et al. [4]. The code of this study uses popular crawling framework OpenWPM [16] and was awarded Distinguished Artifact at the USENIX Security conference. This study measures non-compliance of cookie consent of selected Consent Management Platforms.

Drakonakis et al. [14]. The code of this study is build only using Selenium – lower-level library for interaction with browser than OpenWPM. This study involves registration to websites to measure the security of authentication cookies, and hence is much more stateful than web measurements typically are.

Cassel et al. [6]. The code of this study is written in Java unlike the typical Python of JavaScript crawlers and was awarded Distinguished Artifact at the PETS conference. This study compares privacy among various browsers. For technical reasons, we evaluate only desktop browsers and omit the mobile browsers that are included in the artifact.

Maas et al. [23]. We chose this study as a code-less example. This study measures mis-configuration in IP-annonimization used in Google Analytics, and the simplicity of this task makes it suitable for full replication.

Senol et al. [31] The code of this study uses alternative crawling framework Tracker Radar Collector [33]. This study measures private data exfiltration from web forms prior users submitting the form.

5.1 Reproducibility

In Table 1, we summarize the reproducibility, time needed, along with how the studies satisfied the criteria before and after we embedded them in a container for our Privacy Observatory. Note that we failed to reproduce only one of the works, which happened due to discontinued interface providing access to Google's Single Sign On (*SSO*), on which was the project fundamentally dependent.

As we see in last two columns of Table 1, encapsulation of experiment code into a Docker image greatly improves the satisfaction of Demir's criteria, namely from average of 10.5 to 15. Namely, in Table 2 we report on which criteria are implicitly satisfied using our framework.

In addition to Demir's reproducibility criteria, which are rather theoretical, our practical experience allowed us to state further implementation principles that artifacts need to follow to maintain good reproducibility.

- P1 Limit external dependencies, especially remote ones that can be easily discontinued by the provider.
- P2 Include all binaries. Especially browser binaries are given the fast release cycles unavailable by vendors in short time, and the ever-evolving programming interface of browsers renders binaries of newer versions incompatible.
- P3 Implement proper garbage collection of resources. As crawling can run for long-term, the release of unused limited system resources (memory, number of temporary files, etc.) is crucial. Leaks, might not be evident on local short computation, so testing over extended period is necessary.
- P4 Prevent potential certificate expiry. Many studies measure network traffic between the crawler and internet, for which they require a proxy. This proxy must perform MitM encryption [9], which requires a certificate. If the study cannot utilize alternative measurement methods (Chrome DevTools Protocol [7] or WebDriver BiDi [8]), the certificates should at least be generated dynamically.

5.2 Crawling results

Our results were so far performed only on 100 websites for early testing purposes. These early and limited results include the following observations. Cookie notice compliance is at similar levels as measured in 2021 by Bollinger et al. [4]. Similar to 2021 measurements by Cassel et al. [6], Chrome and Brave browsers are performing fewer third-party requests than Firefox and Tor. The reported levels of Google Analytics IP address anonymizations by Maas et al. [23] are decreasing over time. Similarly the exfiltration of private data from web forms prior their submission is reduced compared to measurements from 2021 by Senol et al. [31]. For more details, refer to Master's thesis by Kast [19].

6 LIMITATIONS

Below, we discuss the constraints and risks that can impact our project.

*Large Language Models (*LLMs*).* Utilizing LLMs, such as GPT, to automate the processing of academic papers, including filtering and systematization, represents a novel approach to the best of our knowledge. Hand-crafted prompts based only on the publication titles have demonstrated only mediocre results, showing that at least abstract is a necessary input. Our strategy involves annotating a substantial dataset with the desired categories, facilitating the computation of performance metrics (accuracy, precision, and recall). This annotated dataset will also enable the training of traditional NLP models, such as BERT [13], serving as both a reference and an alternative in case of high error rate of GPT. In the worst-case scenario, the entire process can be executed manually; however, the primary motivation for automation lies in mitigating the propensity for errors in the labor-intensive and tedious tasks involved.

The exploration of this automation, particularly if successful, could constitute a noteworthy contribution.

Reproducibility issues. The complexity associated with reproducing selected studies can vary widely. As we list in Table 1, we invested 51 hours attempting to reproduce experiments by Drakonakis et al. [14] before ultimately abandoning the effort, but another publication was successfully reproduced within only 16 hours. Still, four of five studies were reproduced successfully, so we remain optimistic about achieving our goal of reproducing 30 publications. Also, the extent to which our measurements may be biased by the use of older software, such as browsers, is unclear. We intend to address this risk by studying how to update the software without compromising compatibility with the existing code. Finally, we acknowledge the bias that might be introduced by our selection of studies or by our capability of reproducing the studies.

Noisy results. Within-study comparisons may yield dubious results due to the inherent noise in web measurements. While we can mitigate noise on our end by repeating measurements, original publications typically do not scrutinize this aspect. A similar challenge was encountered in [5, Sec. 6], and we acknowledge the need for thorough consideration of noise in our findings.

7 CONCLUSIONS AND FUTURE WORK

We created the Privacy Observatory framework to improve the reproducibility of web privacy studies. Successfully reproducing four out of five selected publications validates the design of the Privacy Observatory. Our framework satisfies multiple reproducibility criteria introduced in previous research and new criteria identified by us. The streamlined generation of crawling lists and results processing enables examining results on the same list of URLs across multiple studies. Additionally, the observatory simplifies long-term measurements through configurable repeated deployment, providing a foundation for assessing trends over time. These capabilities offer insights into correlations among various privacy issues and the evolving landscape of web privacy practices.

Looking ahead, we aim to systematize a larger sample of publications. This meta-study and the results of repeated experiments will be a valuable resource for researchers, policymakers, enforcement bodies, and privacy activists. Our ultimate goal is to provide actionable feedback based on the results of both long-term and cross-study measurements, empowering stakeholders to make informed decisions and shape the future of web privacy.

REFERENCES

[1] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021*, WWW '21, page 2165–2176, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Monya Baker. Reproducibility crisis. *Nature*, 533(26):353–66, 2016.

[3] Eleanor Birrell, Jay Rodolitz, Angel Ding, Jenna Lee, Emily McReynolds, Jevan Hutson, and Ada Lerner. Sok: Technical implementation and human impact of internet privacy regulations. *arXiv preprint arXiv:2312.15383*, 2023.

[4] Dino Bollinger, Karel Kubicek, Carlos Cotrini, and David Basin. Automating cookie consent and GDPR violation detection. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2893–2910, Boston, MA, August 2022. USENIX Association.

[5] Ahmed Bouhoula, Karel Kubicek, Amit Zac, Carlos Cotrini, and David Basin. Automated, large-scale analysis of cookie notice compliance. In *33st USENIX Security Symposium (USENIX Security 24)*, Philadelphia, PA, August 2024. USENIX Association.

[6] Darion Cassel, Su-Chin Lin, Alessio Buraggina, William Wang, Andrew Zhang, Lujo Bauer, Hsu-Chun Hsiao, Limin Jia, and Timothy Libert. Omnicrawl: Comprehensive measurement of web tracking with real desktop and mobile browsers. *Proceedings on Privacy Enhancing Technologies*, 2022(1), 2021.

[7] Chromium developers. Chrome DevTools protocol. Standard, International Organization for Standardization, 2015.

[8] W3C committee. WebDriver BiDi. Standard, W3c, 2023.

[9] Aldo Cortesi, Maximilian Hils, Thomas Kriechbaumer, and contributors. mitmproxy: A free and open source interactive HTTPS proxy, 2010–. [Version 10.2].

[10] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. *CoRR*, abs/1808.05096, 2018.

[11] Nurullah Demir, Matteo Große-Kampmann, Tobias Urban, Christian Wressnegger, Thorsten Holz, and Norbert Pohlmann. Reproducibility and replicability of web measurement studies. In *Proceedings of the ACM Web Conference 2022*, pages 533–544, 2022.

[12] Nurullah Demir, Jan Hörnemann, Matteo Große-Kampmann, Tobias Urban, Norbert Pohlmann, Thorsten Holz, and Christian Wressnegger. On the similarity of web measurements under different experimental setups. In *Proceedings of the 2023 ACM on Internet Measurement Conference*, IMC '23, page 356–369, New York, NY, USA, 2023. Association for Computing Machinery.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[14] Kostas Drakonakis, Sotiris Ioannidis, and Jason Polakis. The cookie hunter: Automated black-box auditing for web authentication and authorization flaws. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1953–1970, 2020.

[15] Peter Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies: 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings 10*, pages 1–18, Berlin, Heidelberg, 2010. Springer.

[16] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 1388–1401, New York, NY, USA, 2016. Association for Computing Machinery.

[17] PWC IAB. Internet advertising revenue report: Full-year 2022 results. https://www.iab.com/wp-content/uploads/2023/04/IAB_PwC_Internet_Advertising_Revenue_Report_2022.pdf; Last accessed on: 2024.01.14, 2023.

[18] Georgios Kampanos and Siamak F Shahandashti. Accept all: The landscape of cookie banners in Greece and the UK. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 213–227. Springer, 2021.

[19] Patrice Kast. Privacy observatory: Aggregation system for reproduction of privacy studies. Master's thesis, ETH Zurich, 2023.

[20] Karel Kubicek. *Automated Analysis and Enforcement of Consent Compliance*. PhD thesis, ETH Zurich, 2024.

[21] Karel Kubicek, Jakob Merane, Carlos Cotrini, Alexander Stremitzer, Stefan Bechtold, and David Basin. Checking websites' GDPR consent compliance for marketing emails. *Proceedings on Privacy Enhancing Technologies*, 2022(2):282–303, 2022.

[22] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. Browser fingerprinting: A survey. *ACM Transactions on the Web (TWEB)*, 14(2):1–33, 2020.

[23] Max Maass, Alina Stöver, Henning Pridöhl, Sebastian Bretthauer, Dominik Herrmann, Matthias Hollick, and Indra Spiecker. Effective notification campaigns on the web: A matter of trust, framing, and support. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2489–2506, 2021.

[24] C. Matte, N. Bielova, and C. Santos. Do cookie banners respect my choice? Measuring legal compliance of banners from IAB Europe's Transparency and Consent Framework. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 791–809. IEEE, 2020.

[25] Jonathan R Mayer and John C Mitchell. Third-party web tracking: Policy and technology. In *2012 IEEE symposium on security and privacy*, pages 413–427. IEEE, 2012.

[26] Daniel Olszewski, Allison Lu, Carson Stillman, Kevin Warren, Cole Kitroser, Alejandro Pascual, Divyajyoti Ukirde, Kevin Butler, and Patrick Traynor. "get in researchers; we're measuring reproducibility": A reproducibility study of machine learning papers in tier 1 security conferences. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23, page 3433–3459, New York, NY, USA, 2023. Association for Computing Machinery.

[27] Nathan Reitinger and Michelle L Mazurek. Sok: Considerations in measuring compliance with privacy regulations (research proposal). *7th Workshop on Technology and Consumer Protection (ConPro '23)*, 2023.

[28] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 155–168, San Jose, CA, April 2012. USENIX Association.

Table 2: Demir's reproducibility criteria [11] and how reproducing them using Privacy Observatory framework (*PO*) satisfies them.

	ID	Property	Satisfied?
Dataset	C1	State analyzed sites	●, available in PO
	C2	State analyzed pages	●, available in PO
	C3	State site or page selection	○, typically in paper
	C4	Perform multiple measurements	◐, optional with PO
Crawler build	C5	Name crawling tech.	●, in Docker image
	C6	State adjustments to crawling tech.	●, in Docker image
	C7	Describe extensions to crawling tech.	●, in Docker image
	C8	State bot detection evasion approach	○, typically in paper
	C9	Used crawler is publicly available	◐, upload image
	C10	Mimic user interaction	○, typically in paper
Env.	C12	Describe crawling strategy	●, in Docker image
	C13	Document a crawl's location	●, available in PO
	C14	State browser adjustments	●, in Docker image
	C15	Describe data processing pipeline	●, in Docker image
Evaluation	C16	Make results are openly available	●, available in PO
	C17	Provide a result/success overview	◐, optional with PO
	C18	Limitations	○, typically in paper
	C19	Ethical discussion	○, typically in paper

[29] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. Toppling top lists: Evaluating the accuracy of popular website lists. In *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, page 374–387, New York, NY, USA, 2022. Association for Computing Machinery.

[30] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. "Can I opt out yet?": GDPR and the global illusion of cookie control. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, Asia CCS '19, page 340–351, New York, NY, USA, 2019. Association for Computing Machinery.

[31] Asuman Senol, Gunes Acar, Mathias Humbert, and Frederik Zuiderveen Borgesius. Leaky forms: a study of email and password exfiltration before form submission. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1813–1830, 2022.

[32] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. Clash of the trackers: Measuring the evolution of the online tracking ecosystem. *arXiv preprint arXiv:1907.12860*, 2019.

[33] DuckDuckGo team. Tracker radar collector. https://github.com/duckduckgo/tracker-radar-collector, 2020.

[34] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 4 years of EU cookie law: Results and lessons learned. *Proceedings on Privacy Enhancing Technologies*, 2019(2):126–145, 04 2019.

[35] Anjo Vahldiek-Oberwagner and Jianying Zhou. Statistics of accepted artifacts at Security Research Artifacts given by conference accepted papers at Top Cyber Security Conferences Ranking. https://secartifacts.github.io/ and http://jianying.space/conference-ranking.html; Last accessed on: 2024.01.14, 2023.

A DEMIR'S REPRODUCIBILITY CRITERIA

Table 2 summarizes reproducibility criteria by Demir et al. [11].

Received 9 February 2024