Reproducibility and Replicability of Web Measurement Studies

Nurullah Demir[1,3], Matteo Große-Kampmann[1,2,5], Tobias Urban[1,2], Christian Wressnegger[3], Thorsten Holz[4], and Norbert Pohlmann[1]

[1]Institute for Internet Security—if(is)
[2]secunet Security Networks AG
[3]KASTEL Security Research Labs, Karlsruhe Institute of Technology
[4]CISPA Helmholtz Center for Information Security
[5]Ruhr University Bochum

Abstract

Web measurement studies can shed light on not yet fully understood phenomena and thus are essential for analyzing how the modern Web works. This often requires building new and adjusting existing crawling setups, which has led to a wide variety of analysis tools for different (but related) aspects. If these efforts are not sufficiently documented, the *reproducibility* and *replicability* of the measurements may suffer—two properties that are crucial to sustainable research. In this paper, we survey 117 recent research papers to derive best practices for Web-based measurement studies and specify criteria that need to be met in practice. When applying these criteria to the surveyed papers, we find that the experimental setup and other aspects essential to reproducing and replicating results are often missing. We underline the criticality of this finding by performing a large-scale Web measurement study on 4.5 million pages with 24 different measurement setups to demonstrate the influence of the individual criteria. Our experiments show that slight differences in the experimental setup directly affect the overall results and must be documented accurately and carefully.

1 Introduction

As the Web has grown to an essential part of our day-to-day life, the complexity of the employed web applications has increased drastically. This development has been accompanied by undesirable practices, such as user tracking [57, 19, 32], fingerprinting [47, 21], or even outright malicious activities, such as XSS attacks [63]. Web measurement studies are an essential tool to understand, identify, and quantify such threats, and they allow us to explore isolated phenomenons at a large scale. As the modern Web is highly dynamic and ever-changing, this is an inherently difficult task. To conduct studies across thousands of websites, researchers can partly rely on crawling frameworks such as *OpenWPM* [21], but more often, they have to extend existing work or build new crawlers on their own to adapt to new developments on the Web.

This trend, however, raises the question of whether different measurement studies using different frameworks for gathering data are comparable *and* to which extent experiments can be reproduced or replicated. In particular, in the field of Web-based measurements, ensuring replicability requires a tremendous effort to describe, document, and openly communicate the details of the experimental setup and implementations. However, if the community cannot verify and reenact drawn conclusions, the entire scientific process is at risk of becoming unreliable—something that has unfortunately been observed in different research disciplines in the past [31, 54, 17, 44].

In this work, we systematize such effects, provide best practices and criteria that help design studies, and additionally perform a large-scale Web measurement study that highlights the impact of these subtle differences. In particular, we survey

117 research papers published at top-tier security and privacy venues in the past six years. Based on this survey, we factor out common fundamental principles for Web measurements and establish common guidelines for conducting such experiments. We define criteria that help designing experimental setups that are reproducible and replicable. By applying these criteria to the analyzed papers, we find that the documentation of the experimental setups is often neglected and does not fulfill the community's expectations of a Web measurement study (see Section 4). In a large-scale study for which we visit 4.5 million pages on over 8,800 sites with 24 browser profiles, we show that slight changes in the experimental setup alters the results to an extent where cross-comparability of studies is not feasible (see Section 4). For example, we find that the identified trackers on pages can vary by 25% based on the used browser configuration.

In summary, we make the following contributions:

- **Guidelines for Web measurements.** We highlight the challenges of designing Web measurements and provide guidelines that help setting up experiments that effectively address them.
- **Prevalence study.** We perform a survey of 117 security and privacy papers from 2016–2021 that perform Web measurements and show that our described challenges affect most of them.
- **Impact analysis.** To increase the comparability of future and previous Web measurements, we perform experiments utilizing 24 measurement setups and compare the measured differences that emerge from the utilized frameworks.

2 Designing Web Measurement Studies

The rapidly changing, variable content and the general trend towards providing more content online makes the Web a challenging subject for measurement studies. As an example, suppose one visits the same website at the same time with different browser instances. The loaded content (e.g., ads or other dynamic content) likely differs and, thus, the overall results of measurement studies might deviate (e.g., when identifying embedded trackers or analyzing shown ads). This simple example illustrates that repeated experiments may show (slightly) different results and conducting such experiments in an uncontrolled environment is bristled with obstacles, endangering replicability. However, the cornerstone of academic work is the possibility to scrutinize conclusions and results. We thus pick up the definitions of the Association for Computing Machinery [9] for a) repeatability (*"Same team, same experimental setup"*), b) reproducibility (*"Different team, same experimental setup"*), and c) replicability (*"Different team, different experimental setup"*).

We put a particular emphasis on reproducibility (see Section 3) and replicability (see Section 4) of published studies, and leave repeatability aside, as by definition it can only be achieved by the team that conducted the experiment in the first place. Thus, reproducibility and replicability are essential to our analysis, as these enable us to verify and compare results of existing work.

2.1 Literature Survey

Transparency is an essential factor in producing reproducible and replicable experiments. To understand the current state-of-the-art of Web measurements in the security and privacy community, we perform an extensive literature survey based on publications at the top-tier conferences in this community: *IEEE S&P*, *ACM CCS*, *USENIX Security*, *NDSS*, *PETS*, and the "Security, Privacy, and Trust" tracks at *ACM TheWebConf* as well as *ACM IMC*. We performed the survey across the past six years (2016–2021).

2.1.1 Paper-Selection Criteria

Of course, not all papers on the surveyed venues conduct a Web measurement or rely on data collected by such a study. Therefore, we first determine the papers of interest based on the following characteristics: (1) The paper in question analyzes a phenomenon present on websites (e.g., embedded third parties or used libraries) or focuses on the communication with a website (e.g., HTTP headers), and (2) the paper in question analyzes more than one website. The definition allows us to focus on works that, on the one hand, study similar research objects (i.e., websites and their communication) and, on the other hand, need to scale their experiments comparably. In a first step, we analyze 4,407 papers from the above-mentioned venues and determine

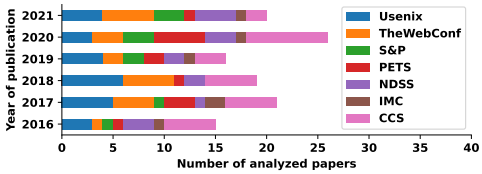

Figure 1: Number of surveyed papers that perform a security or privacy Web measurement, per year and venue.

whether or not to include them in our survey, by skimming the title, abstract, and method. From the entire corpus, we consider 117 (2.7%) papers to be analyzed in depth. Fig. 1 details the number of surveyed papers per year and venue. Of those papers, 71 (61.5%) focus only on security challenges, 35 (29.9%) on privacy issues, and 11 (8.5%) on both. This general overview of our survey shows that Web measurement studies are an important tool in the security and privacy community to analyze different phenomena and push the field forward. It hence is essential to investigate how our community performs studies, derive best practices, and analyze to which extent existing studies allow to reproduce the results of the experiments.

2.2 Challenges and Best Practices

In the following, we present design patterns and best practices that help to plan Web measurement studies so that future studies can be designed to be reproducible. We create these guidelines based on the surveyed literature and our own experience in this research area. For all surveyed papers, we analyze the documented setup of each experiment, abstract general design choices, and develop best practices that are intended to provide an overview of aspects which need to be considered when designing Web measurements in practice. It is essential to highlight that our guidelines are not intended to point fingers or criticize previous work, but to highlight pitfalls and challenges that can impact a study's outcome.

2.2.1 Method to Design the Best Practices

To derive the best practices, we analyze different experiment design choices of the papers and compare the outcomes of the works. This allows us to identify generalizable and common aspects that are shared across different works. For example, if one work visits sub-pages and another work only visits landing pages but both find different levels of tracking activities, we create a best practices that researchers should take this behavior into account. Moreover, we use these best practices to derive criteria that measurement studies should follow to allow for reproducibility of conducted experiments.

2.2.2 Building the Dataset

Naturally, each Web measurement study has to identify websites and pages to be analyzed during the experiment. For this step, one can distinguish between three methodically different approaches, which all come with up- and downsides.

P1 Artificially selecting websites and pages. As the Web is ever-growing and consists of a myriad of sites with even more pages, measuring all of them in a single experiment is not feasible in a reasonable way. A commonly accepted way of focusing an experiment is to use a so-called "top-list" that ranks popular sites (e.g., *Alexa* [5], *Tranco* [38], or others [7, 40]). These lists, however, only include the landing page (or the eTLD+1) which are used for the experiment. While, at first glance, this might seem reasonable, recent works have shown that subsites (e.g., *https://www.example.com/news*) show a significantly different behavior than the respective landing pages [56, 7], and that the rank of a website also might impact the results [64]. Consequently, we advocate to name the sources (e.g., top-list) of sites that have been analyzed, detail how they have been picked, and list all analyzed pages (e.g., in an appendix). Similar to the highlighted challenges to enable repeatability, it is necessary to point out which criteria are used to choose or eliminate entries from a given set of sites.

P2 Using user clickstream data. Another approach is to use clickstreams observed from real users or analyze their traffic directly. While such an approach is more realistic, they are harder to collect. However, studies that explicitly need to understand the effects of a phenomenon for individual users need to take this step [11, 46, 22]. If 'only' the presence of a phenomenon is of interest (e.g., if secure CSPs are used), artificially selected sites can suit the purpose.

P3 Use existing sources. Using a previously collected public datasets that contains measured

Web traffic (e.g., *HTTPArchive* [27]), is the only option that allows the reproduction of results, offers high repeatability, and enables to compare properties. However, one is bound to analyze phenomena for which data is already present in the desired granularity [18], which is often not the case.

From this set of best practices, we derive four criteria (C1–C4) that a measurement study should meet (see group *"Dataset"* in Table 1).While criteria C1–C3 (*"documentation of the analyzed sites"*) are directly related to the named practices, criterion C4 is intended to highlight that some phenomena need to be analyzed over time to understand their scale.In the surveyed papers, C4 was often not noted and we analyze its effects in Section 4 in detail.

2.2.3 Experiment Design

One way or another, Web measurement studies rely on a crawler.Selecting, building, and customizing such a crawler is an essential step in preparing each study, such that one needs to prudently design and implement it to ensure that the experiment is stable, repeatable, and comparable.

Building the Crawler We now discuss design decisions when performing a study using artificial browsing data (i.e., not using user-generated or public data). We review the essential steps that should be taken into account when designing such a study:

P4 Choosing a technology. Previous work relies on different measurement setups ranging from simple tools like *cURL* [15] to sophisticated scalable measurement frameworks that can spawn several browsers at once like *OpenWPM* [43, 21]. As prior work has shown, the decision of which tool to use impacts the results [4].

P5 Customization of the crawler. Naturally, each study uses a (slightly) different measurement setup. When customizing a crawler, it is inevitable to elaborate on the steps taken and discuss possible artifacts and limitations of the endured approach. While necessary, each customization step might impact the results (e.g., using different user agents) and, therefore, needs to be documented [39]. We discuss these effects in more detail in Section 4.

P6 Avoiding crawler detection. Crawlers and other bots make up approx. 37% of traffic on the Web [29] and it has been shown that this significantly affects crawling studies [62, 30, 41]. Consequently, some service providers define behavior guidelines to limit crawling traffic, or try to detect and block them altogether [33]. These defense mechanisms might substantially impact the results of measurement studies if sites present different content or none at all. Hence, the authors' choice to avoid and if so how and to which extent an evasion technique was implemented needs to be discussed transparently. However, it is commonly accepted (and often necessary) to circumvent bot detection mechanisms [56, 57, 21].

P7 Mimicking User Interaction. Modern websites are no longer static HTML pages, but interactive applications that load different sets of content depending on the users' actions. Resources are often only loaded once visible to the user (known as "lazy loading") to improve the website's loading speed and for search engine optimization purposes [25]. This means that crawlers that do not interact with a page (e.g., scrolling) will miss crucial resources [56, 57, 34, 65]. Therefore, interaction mechanisms need to be documented, and limitations of lacking user interaction should be discussed.

Based on these four aspects of a crawling setup, we derive criteria C5, C6, C7, C8, and C10. We split the customization step (P5) into two criteria (C6 & C7) to account for differences whether a crawler was modified (e.g., a function was altered) or extended (e.g., a browser extension was used). Furthermore, we add a criterion that urges authors to make the crawler publicly available (C9). Since the effects of C5 and C11 are not yet adequately discussed by previous work, we analyze them in Section 4.

Experiment Environment After selecting the sites to visit and building the crawler, the experimental environment must be crafted. In the following, we describe essential environmental aspects that may impact the crawler and, therefore, the experiment's outcome.

P8 Geolocation of crawls. A critical factor for each experiment is the location from which the measurement study is conducted. Depending on the location (e.g., based on the IP address of the crawling machine), websites might deliver different content [28]. This may, for instance, be founded in cloaking, legislation (e.g., the GDPR or

CCPA [64, 57, 16, 56]), or even censorship [45, 12]. Such impacts have to be accounted for (e.g., via using a VPN setup) and actions to address them need to be disclosed in detail.

P9 Defining the page visit strategy. For the page visit strategy, we distinguish between *stateless* and *stateful* crawls. A stateless crawler (i.e., browser) is reset completely between each page visit, such that each visit creates a new HTTP session that updates the browser's internal resources. On the contrary, some (e.g., only the cookie jar) or all of this information is kept in stateful crawls, as a "real" browser would. Consequently, authors need to document what part of a browser profile is maintained statefully, what part is reset, and when [21, 65]. This distinction has a severe impact on the outcome of the experiment: In stateful experiments, the order of visited pages potentially impacts the results, and it accounts for HTTP session-specific phenomena, such as opt-in to cookie tracking. Stateless crawls, in turn, allow to study session-independent attributes. Note that this practice does *not* account for browser profiles that were populated before the measurement took place (e.g., by pre-filling the cookie jar), we account for this in P11.

P10 Setting up Browser Configuration. A browser' configuration plays an important role for Web-based measurements. Depending on the browser (e.g., version) the crawled entities might act differently. To allow comparability and reproducibility of experiments, it is essential to share basic configuration details, which may impact the study's outcomes [37, 60, 65]. Such design choices range from installed extensions, used block-lists, login strategies, used browser version, content of the cookie jar, etcetera.

P11 Describe shortcomings and limitations. Naturally, a Web measurement can never be complete regarding, for instance, coverage or realism. The experimental design accounts for these "natural" boundaries, but each design choice will likely impose certain restrictions and limitations. To allow the research community to acknowledge and assess the outcomes of an experiment fully, it is inevitable to discuss the limitations of its design [52, 50].

From these practices, each can be mapped to a single criterion (C11, C12, C13, and C17). We add an additional criterion (C15) asking to make results publicly available, as this particularly helps to replicate or reproduce an experiment. Moreover, we set up two criteria that help assessing the findings of a paper: First, asking for an ethical discussion (C18) and second, urging to provide a general overview of the measured results (C16).

2.3 Design and Evaluation Criteria

Based on the best practices described in the previous section, we derived the named 18 criteria to allow reproducibility of a study. In a first step, two experts, both with an extensive professional and academic background in security and privacy on the Web, assessed an identical, randomly selected subset of the surveyed papers ($n = 25$) to test the applicability of the criteria. This exploratory evaluation has shown a very high interrater reliability (Cohen's kappa: $\kappa = 0.94$), which indicates that the designed criteria can be unambiguously applied. In a few cases, the experts have disagreed, which however turned out to be founded in an initially ambiguous formulation of one criteria, which was adjusted accordingly. In a second step, the criteria have then been applied to all 117 papers in our corpus. Table 1 lists all 18 criteria and provides a brief description of each.

3 Evaluating Reproducibility

In this section, we analyze the surveyed papers along with the criteria we have introduced to get an understanding of the reproducibility of previous works. The decision if a criterion is (fully) satisfied is not always binary. For example, a paper might state that the crawler was instrumented but omit how.

Evaluation Categories We use the following four categories to distinguish if and how a criterion is satisfied:

- **N/A:** The criterion does not apply to the analyzed paper as it does not impact the used methodology. For example, **(author?)** [38] crawl four toplists and combine the results in a sophisticated manner. In this case, the criteria *"Mimic user interaction"* (C10) or *"Geolocation"* (C12) do not apply.
- **Omit:** A paper does not state the taken actions to satisfy a criterion, but it would be essential to

Table 1: Criteria to design Web measurement studies.

		ID	Criterion	Description
Dataset		C1	State analyzed sites	States used dataset, toplist, or user clickstreams, including version.
		C2	State analyzed pages	Offers a `.csv` or comparable with all analyzed pages (i.e., distinct URLs).
		C3	State site or page selection	Discusses the selection process of analyzed sites.
		C4	Perform multiple measurements	Discuss which pages are analyzed in consecutive measurement runs, if appropriate.
Experiment Design	*Building the Crawler*	C5	Name crawling tech.	Describes the used crawling technology (e.g., *OpenWPM*).
		C6	State adjustments to crawling tech.	States which technology features were used and/or (slightly) adjusted.
		C7	Describe extensions to crawling tech.	Describes which *new* features were developed to conduct, if any were made.
		C8	State bot detection evasion approach	Discusses which means were taken that the crawler was not detected, if necessary.
		C9	Used crawler is publicly available	Provides the crawler in a public location.
		C10	Mimic user interaction	Describes how the user interaction was implemented, if applicable.
	Experiment Env.	C11	Describe crawling strategy	Describes which crawling strategy was used (e.g., stateless vs. stateful).
		C12	Document a crawl's location	States from which location(s) the study was conducted.
		C13	State browser adjustments	Discusses properties of the browser (e.g., user agent, version, used extensions).
		C14	Describe data processing pipeline	Describes the data processing steps in detail.
Evaluation		C15	Make results are openly available	Authors provide the (raw) measurement results.
		C16	Provide a result/success overview	Describes the outcome of the measurement process on a higher level.
		C17	Limitations	Discusses the limitations of the experiment.
		C18	Ethical discussion	Discusses ethical implications of the experiment (e.g., exploiting vulnerabilities).

reproduce the work or that it potentially affects the outcome of the work.

• **Undocumented:** If a paper states that the authors took actions to satisfy a criterion but do not specify how. For instance, the authors state that "measures were taken to avoid bot detection" but do not explain how this has been implemented.

• **Satisfies:** This is the desirable case in which a paper satisfies a criterion and details which measures have been taken to do so.

These categories allow us to differentiate to which extent a criterion is satisfied and enable us to perform a fine-grained analysis of the reproducibility. Note that these categories are *not* meant to indicate whether the taken actions in a paper are sound or complete to satisfy a criterion. Rather, they aim to understand if and to which extent an experimental setup can be rebuilt

3.1 Survey Results

We analyze all 117 papers from our survey (see Section 2.1). Across all categories and papers, merely in 33 (1.6%) cases a criterion does not apply to an analyzed paper at all (category *N/A*). Criterion C16 (*"General result/success overview"*) is satisfied by most papers (115 (98.3%). Relating to all criteria and papers, more than two-fifths of all criteria are satisfied (882 (41.9%)), in 1,055 (50%) of all cases the paper omits any information on the criteria, and in 136 (6.5%) cases a criterion applies to a paper but the paper does not include a description of it (category *Undocumented*).In Appendix B, we conduct analyses of each criterion individually.

3.1.1 Dataset

We only found twelve papers (10.5%) that fulfill all four criteria related to the dataset. However, 64.1% of the papers state the dataset they used. Four (3.4%) of the papers do not state which sites they analyzed. Furthermore, the vast majority (72.6%) does not offer a complete list of all the analyzed pages. Regarding the reproducibility of the experiments, these results are critical because most experiments are not reproducible regarding the sites and pages that have been analyzed. The papers that used a Tranco list [38] all offer a list of visited sites, which shows that works that aim to provide best practices have a positive impact on our community. Another result is that 63.3% of the analyzed papers do not perform measurements in multiple measurement runs. Conducting a measurement only once might offer little insight into generalizability, as the experiments of Agarwal et al. indicate [3, 2].

3.1.2 Experimental Design

Three of the five criteria (C7, C8, and C9) in this category are omitted by at least half of the analyzed papers. While most papers state the crawler, many fail to address whether configurations have been changed or extensions have been used. This result is concerning because documenting adjustments to the crawling technology is an essential part of understanding and rebuilding an experimental setup. Most of the papers crawl data from websites but do not state how they evade bot detection or make the crawler publicly available, which raises transparency and ethical issues. Our analysis indicates that approximately a third (30.7%) of the papers submitted to the top measurement, security and privacy conferences are not stating which technology they used to crawl. This result again is having a severe impact on the reproducibility of the experiments, as these design choices potentially have significant impact on the results [4, 34]. Criteria C10, C12, and C13 are omitted by 69.5% of the analyzed papers on average. For C10 and C13, the omission might be due to the fact that recently it was systematically shown that these factors play an important role [56]. C12 is omitted by more than two out of three papers (71.8%) that do not state where the scan is geographically located. In our analysis in Section 4, we show that this can significantly impact the overall reliability of the results and the reproducibility of the experiments. However, 76.1% of the analyzed papers describe their data processing pipeline, such that it becomes clear how the crawled data is processed for analysis. However, approximately 17.1% of the papers, where the pipeline is described, fail to offer details on the crawling technology, making the reproducibility of the analysis impossible. Combining criteria from the experimental design (C8 + C10 + C11 + C12), we can deduct an analysis about the realism of the papers. Except for C11 (omitted by 41% of the works) more than half of these criteria are omitted by the papers.

3.1.3 Evaluation

We have not observed a single paper where the evaluation is not applicable. However, we find that more than half (64.1%) of the analyzed papers omit an ethical discussion. This is questionable in the discovery and detection of vulnerabilities. Research that measures these on a large scale should always include an ethics section. Roughly, 21.3% of the analyzed papers miss a limitations *and* ethics section, which can be considered as a disputable research practice. In terms of open science, only 24% of the analyzed papers make their results openly available.

3.1.4 Venue Comparison

To understand venue-based differences, we cross-compare papers from the analyzed venues and nine essential criteria. For this analysis, we only consider criteria that must be met to allow the repetition of an experiment. These criteria are: C1–C3, C5–C7, C11, and C12. We do not see a tendency that any venue publishes works that describe the methodology approach better or worse than other venues. The only exception is *ACM IMC*. The criteria were omitted 9 (18.7%) times. The other conferences have an average omit rate of 56.8% with a standard deviation of 46.7%. More than half of the papers, except for *IMC* and *PETS*, neglect the geolocation of the crawls. *USENIX Security* is the only conference where six criteria are omitted by more than half of the papers. This conference is the only conference that utilizes an artifact evaluation. Unfortunately, only one of the analyzed papers received such a batch and, therefore, we cannot generalize the usefulness yet. Especially the crawling criteria (C5–C7) and location of the measurement origin suffer from violations at all top-tier conferences. While the absolute numbers and ratios of criteria violations are generally comparable, we conclude that any Web measurement research published can equally benefit from the criteria we defined in the paper. The *PETS* symposium is the only conference in our corpus where not a single paper conducted an ethical review of their work. This is an at least unsettling finding, because privacy and ethics are intertwined and must be taken into account when conducting privacy measurements [26].

Self-Reflection This work focuses on reproducibility and replicability of Web measurement studies and highlights the need for proper documentation and provision of needed supplementary material. However, in line with similar works [53, 8, 58], we chose not to publish the raw results of the categorization process. It is not our intention to blame individual works for flaws—for which our own papers are no exception—but to raise awareness for a general potential problem in our community.

4 Case Studies

In this section, we proceed to demonstrate the impact of insufficiently documented experimental setups of large-scale studies along four exemplarily case studies focusing on C4, C5, C10, and C12. The first three are chosen because the literature currently does not provide enough evidence on their impact, while C12 is used to verify that our framework is able to reproduce previous results.

4.1 Web Measurement Approach

To show the impact of seemingly small changes in a measurement setup on the replicability of an experiment, we perform a Web measurement study that uses 25 different setups. More specifically, we compare results of four browsers (*Firefox*, *Firefox headless*, *Chrome*, and *Chrome headless* – C5), three regions (Europe, Asia, and North America – C12), and two types of website interactions (*"none"* and *"simple interaction"* – C10) individually. Overall, we cross-compare 24 different setups. Additionally, we perform a repeating study that measures the same sites and pages on a daily basis – C4.

The data corpus of our study consists of the top 10k Tranco [38] websites and we collect the first 25 subpages for each site (as identified by the JavaScript engine), if possible. We designed a pipeline that coordinates all page visits across the profiles. Our measurement setup consists of virtual machines (VMs) orchestrated by a "commander" instance to organize parallel page visits. For example, one VM performs the measurement using Chrome from the US with user interaction, while another does the same for the EU. A detailed description of our framework can be found in Appendix C.

The commander is in charge of starting the measurement for each site in parallel across the VMs. Each VM starts 10 browsers (one for each site) in parallel using the defined profile. Once the analysis of a page is finished, the same browser instance is moving to the next page of the same site. Hence, subsequent visits of pages will not be synchronized across all VMs. On the landing page level, the timing differences in our experiment are only 17 seconds on average. However, the timing differences on subpage level are 3 min (SD: 7 min). When visiting a page, each browser logs all HTTP requests and responses and stores them in a central database. We wait until a page has finished loading or a timeout of 30 seconds is reached, close the browser, and move on to the next page.

4.2 Replicability of Measurements

We highlight the impact of individual criteria based on four examples and explore them along two dimensions: (1) Web tracking and (2) usage of *Content Security Policies* by a page.

4.2.1 Method

To compare the results of the 24 profiles, we use the *Jaccard index*.For each page, we have a set of observed trackers and CSPs (i.e., 24 sets). The Jaccard index is used to gauge the similarity of sets. The index computes the similarity by dividing the size of the intersections with the size of the union of all sets. By design, the index ranges from 0 to 1, where 1 denotes that the sets are equal and 0 indicates that they have no element in common. This allows us to compare and quantify the differences in observed trackers on page level across all profiles. The Jaccard index is used to compare the similarity of two sets. Since we compare multiple sets, we compute the pairwise similarity between all sets and use the arithmetic mean to state the similarity for a given page.

We analyze the impact on privacy-related studies along with the presence of trackers. More specifically, we analyze which tracking requests are observable when visiting a page. To identify them, we use the tracking filter list *EasyList* (as off 07/05/2021) [20]), which we provide in the supplementary data of this work (see Section A). If an observed URL is present on the list, we consider it to be a tracking request. Furthermore, we use the eTLD+1 part of these URLs to identify *trackers*, in terms of domain names.

To get a better understanding of the impact of different measurement setups for security studies, we analyze the presence of *Content Security Policies* (CSP). CSPs help to mitigate specific attack vectors on the Web (e.g., XSS attacks). They are implemented by an HTTP header that contains different directives that define sources from which content may be loaded. We analyze differences in CSPs by inspecting the used CSP directives and all attributes within a directive. We omit all variable attributes (e.g., nonces) in the analysis since they change by design. Furthermore, we compare the semantic effect of a directive (i.e., ordering is ignored).

4.2.2 General Measurement Overview

Our total website corpus consists of 10k distinct sites and we found 182,586 subpages on those sites, including the landing pages. Across all profiles, we successfully visited 4.5M pages on 8,883 sites. The sites that could not be crawled are not meant to be visited by a human (e.g., link shorteners, content delivery networks, or ad networks). The resulting database has a size of roughly 1.1 TB, which is openly available (see Section A). On average, each profile visited 179,404 pages (SD: 6,947; max: 186,972; min: 158,691). In our analysis, we only consider pages for which we observed at least 17 successful crawls across the 24 profiles. Hence, roughly 70% of the profiles have to visit a page so that we consider it. Furthermore, this guarantees that at least one profile in each category successfully crawled a page. 178,452 (92%) of the analyzed pages fall into this category. Note that 134,120 (75%) of pages were successfully crawled by all profiles.

Figure 2 provides an overview of the number of observed tracking requests and trackers (eTLD+1) for each page by profile. Generally, we see that *Firefox* profiles are tracked more than their *Chrome* counterparts. Furthermore, profiles in the US are tracked more than profiles from other regions. Finally, user interaction seems to have a significant effect in terms of tracking, while running browsers in headless mode makes only little difference. More details about our results are presented in the following sections.

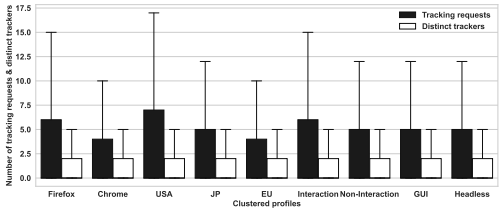

Figure 2: Observed tracking requests and trackers by profile.

4.2.3 Impact of Different Browsers (C5)

First, we study the impact of the four browsers that we analyzed (i.e., *Firefox*, *Firefox headless*, *Chrome*, and *Chrome headless*) in terms of their impact on tracking. Regarding HTTP requests, we see that *Chrome*-based profiles make on average 2% (SD: 18.5%) more HTTP requests than *Firefox*-based profiles. Furthermore, we see that every 10th (SD: 1.5, min: 7, max: 12) HTTP request is a tracking request. We observed that for *Chrome*, every 10th HTTP request is a tracking request, while for *Firefox* it is every 9th. Overall, we identified for all Firefox profiles 12% more tracking requests than for the *Chrome* profiles. Only for four out of the 12 *Chrome* profiles, we could detect more trackers than for the respective *Firefox* profiles. However, on average, we identified 3.9 (SD:8.6) distinct trackers (eTLD+1) per page for the *Firefox* profiles and 3.9 (SD:8.1, min: 0, max: 68) distinct trackers for the *Chrome* profiles.Hence, the number of distinct tracking domains stays similar, while the volume of requests differs between the two browsers.

We turn to the effects of when a browser is used in headless or native ("GUI") mode. We observed only in two of the six *Chrome headless* profiles more trackers (10% per page) than in their counterparts. Overall, the headless *Chrome* profiles only contained 3% fewer trackers. When we run *Firefox* in headless mode, we noted almost reversed results. For four out of six *Firefox Headless* profiles, we could detect 5% more trackers than native *Firefox* profiles. Across all of these profiles, we see marginal differences regarding the number of trackers when we run browsers in headless mode. This is in contrast to previous work that has shown the importance of this feature [4]. However, in some of the profiles, we observe substantial differences, which indicates that the outcome of an experiment is not determined by the used browser mode exclusively. Moreover, different combinations of design choices mutually affect the results, highlighting the need for proper documentation.

Across all pages, the mean Jaccard similarity in observed distinct trackers for browsers *Chrome* and *Firefox* is 0.59 (SD: 0.32, min: 0, max: 1). Overall, we identified only 1% more trackers for *Firefox headless* profiles. However, we find a big difference in terms of identified trackers. Across all pages, the mean Jaccard similarity in observed distinct trackers for headless and non-headless profiles is 0.53 (SD: 0.48, min: 0, max: 1). Overall, the similarity in observed trackers comes with a medium Jaccard similarity for this category but with a significant standard deviation. While we observed a perfect similarity (1) for 19% of the pages, we found no similarity (0) for 11% of them (see also 3). This effect is magnified if we only look at the headless and non-headless browser where we find perfect similarity for 35% of the pages and no similarity for 34%. Hence, in the worst case, studies that only alter the browser (or the display mode) might find different results, depending on the analyzed pages.

The distribution of the computed Jaccard values for each page is given in Figure 3 (black bar). Most pages (34.1%) always issue a very similar set of trackers no matter which profile visited the page (similarity ≥ 0.8). It is worth noting that we identified on such pages 1.9 distinct trackers on average. These pages only contain few trackers, but those are often present independently of the used profile. On 45.5% of the analyzed pages, we found a medium similarity ($0.3 \leq$ sim. < 0.8) in the observed trackers. On those pages, we observed on average 5.2 trackers. Finally, 20.4% of the analyzed pages show almost no similarity (< 0.3) in the observed trackers. On those pages, we observed on average 4.0 trackers. Thus, pages that include more trackers also include a different set of trackers based on the used profile. In the following sections, we discuss the impact of other criteria on the similarity in more detail.

Our results show that the number of tracking requests observed in the *Firefox*-based measurements is higher than in the *Chrome*-based ones. However, we did not find a statistically significant effect that running browsers in headless mode affects the number of observed trackers. However, we find a statistically significant difference (p-value < 0.001) in terms of identified distinct trackers.

We now describe our security analysis regarding CSP. Overall, we identified CSPs on 17.596 pages

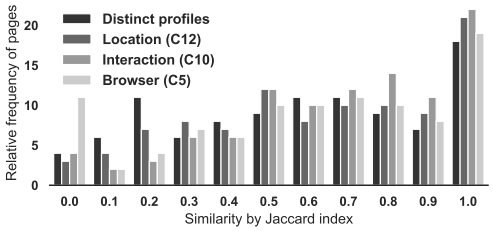

Figure 3: Similarity of trackers on page level by profiles.

(10%). Compared to tracking analysis, we find very high similarity for CSPs. We find that on 16.355 pages (93%) the identified CSP headers are semantically identical. Hence, overall we get a Jaccard similarity of roughly .97. However, on the pages that served different CSPs (1.063), the mean Jaccard similarity is 0.68 (SD:0.25). Furthermore, we did not find that any profile had a significant impact on this phenomenon. This result is expectable since some of our features cannot be detected when the website is visited (e.g., user interaction) and, thus, cannot impact the results. Due to the low impact of our profiles on served CPSs, we dropped the CSP analysis in the other section because we found comparable results. Future work could analyze the impact of different browser profiles on more variable security features.

4.2.4 Impact of Simulated User Interaction (C10)

Regarding simulating user interactions, our analysis shows that interaction on pages causes a sharp increase in HTTP traffic (on average by 20%) while the number of tracking requests increases by 35%. Hence, the amount of tracking requests increases disproportional with the number of all observed requests. For profiles with interaction, we observed on average 7.2 (SD: 8.8) distinct trackers (eTLD+1) per page and for the other profiles 6.7 (SD: 8.3). Thus, these high-level figures already indicate that the choice to simulate user interaction impacts the results of a study. When analyzing *Chrome* and *Firefox* separately, we see statistically significant (p-value < 0.001) differences. This again indicates that the effect of a single criterion cannot straightforwardly be attributed but that they jointly impact the results. For *Chrome* we find on average 6% (SD: 10%, min: -9%, max: 14%)) more HTTP requests when we perform interactions and, surprisingly, we see an average increase of 36% (SD: 6%, min: 29%, max: 43%) for the *Firefox* profiles. Of these requests we see that for *Chrome* 5.6% are tracking requests. For *Firefox* we see that 73% (SD: 21%, min: 43%, max: 92%) of these request are used to track users. This difference might be an artifact of our measurement framework and should be analyzed in future work in more detail.

Across all pages, the mean Jaccard similarity in observed distinct trackers for profiles with *interaction* and *non-interaction* is 0.67 (SD: 0.28 min: 0, max: 1). These results fit the observation that the number of observed trackers (eTLD+1) does not increase by a lot by user interaction. If the number of trackers stays similar, one can expect that the set of trackers per page stays similar. Almost half of all pages (47%) show a high similarity of more than 0.8 (see also Figure 3), which also indicates this trend. However, for a third (33%) of all pages, we find a similarity of 0.5 or less. This shows that while the overall similarity is quite substantial for a non-negligible number of pages, the results differ considerably.

4.2.5 Impact of Different Locations (C12)

In this section, we want to analyze the regional effects of an experiment. On average, we see that profiles from the USA are tracked most in terms of distinct trackers (eTLD+1) on a page (6.93; SD: 8.5), followed by Japan (5.6; SD: 6.39), and EU profiles (4.49; SD: 5.42). These results propagate to the number of observed tracking requests.

Across all pages, the mean Jaccard similarity in observed distinct trackers for the profiles in the different regions is 0.62 (SD: 0.30 min: 0, max: 1). In terms of the analyzed criterion, the location has a significant effect on tracking and has only a limited effect on the difference in observed trackers. Half of all analyzed pages (50%) show a similarity of 0.7 or more (see Figure 3), and only 29% of the pages show a similarity of 0.4 or less—which should be accounted for in an experiment. This is in line with previous work that found that only a few online advertising companies altered their business model due to privacy regulations (e.g., withdrawing from the European market) [57]. Overall, we note that privacy measurements and analyses can vary up to 65% depending on the region (e.g., due to different

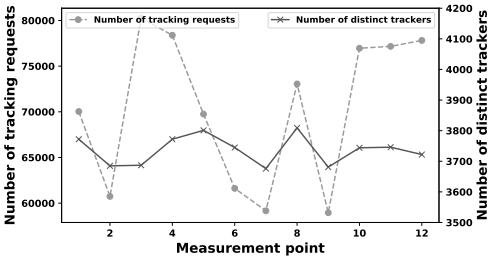

Figure 4: Fluctuation in the number of tracking requests and domains for each day in the long running experiment.

legislation). Thus, our analysis finds that the region plays a crucial role in privacy measurements. These findings are consistent with previous work [57, 56].

4.2.6 Impact of Multiple Measurements (C4)

Finally, we want to assess the temporal effects for measurement studies. In the following, we look at the absolute number of identified tracking requests and the number of observed distinct tracking domains (eTLD+1). Fig. 4 shows the result of this analysis. Over twelve days, we saw a variation of up to 27% (max on day 3–80,274; min on day 9–58,951) in observed tracking requests. The standard deviation of such requests is 8,203. However, the number of distinct tracking domains remains almost stable during the experiment (variation of 3.5%). Our results suggest that depending on the day of each measurement, the number of tracking attempts—in terms of tracking request—frequently varies, but the companies (domains) that are active in the ecosystem remain stable. Thus, studies that analyze the ecosystem will find similar results, while studies that aim to analyze the extent of a tracking phenomenon might see different results based on the measurement day. In terms of replicability and reproducibility, this is challenging since even the same setup measures different levels of tracking on different days, which might lead to different conclusions of a study. Our results show how important repeated measurements are to draw more robust conclusions.

5 Related Work

Recently different works, similar to our measurement study, focused on the comparability of various crawling tools used in Web measurement studies. Most recently, **(author?)** [34] compared how different measurement tools and setups affect the results of a study. Similarly to us, they used other locations and browser modes to perform the measurement. In their study, the authors focus on "request/traffic volumes", JavaScript libraries loaded, and known ad/tracking domains loaded. **(author?)** [4] presented a survey on tools used in Web measurements and performed an experiment to compare the outcomes of these tools. In their study, the authors compare metrics like request/response sizes or used cipher suits. Both works find, similar to our results, that different crawlers impact the results of an experiment. Cassel et al. found that mobile browsers receive fewer tracking-and-advertising requests than desktop browsers in a comparative study [14]. In contrast to our work, they focus on tracking and show differences between mobile and stationary devices. Similar to our approach to systematize and evaluate our community's researcher methods, other studies were performed by various authors in different domains [53, 6, 48, 8, 36]. Our work focuses on a different research object than the named studies, namely the Web.

6 Ethics & Limitations

One limitation of our measurement is that—for scaling reasons—the site visits are not fully synchronized. We argue that this limitation will have minor influence on the results of our study, as the site visits still happen within a small time window (mean time difference is 3 min). We thus assume that sites will still hold similar content. The experiment design comes with the limitation that our crawler does not interact with websites as an actual human would, which is probably impossible in an automated fashion.From an ethical point of view, our crawler creates traffic on the visited websites that could be omitted and save resources, and we might see ads that might drain the budget of the advertising company. Since our crawler only visits each page once (once a day for the long-running experiment), we argue that these issues are minor

and can be accepted.Our survey comes with limitations. There are more conferences than the analyzed venues. However, we decided to focus on them because (1) works published there go through a very competitive process, which is meant to increase the quality of the published work, and (2) we argue that analyzing 117 papers provides a sample set that is large enough to make qualified assertions.

7 Conclusion

Our survey shows that Web-based measurement studies often do not sufficiently document their experimental setup. As an example, the used crawler or its configuration are frequently not described in detail. This results in a lack of *reproducibility* in practice. To help mitigate this in the future, we have developed a set of best practices and 18 criteria for designing and conducting Web measurement studies. In a large-scale measurement with multiple crawling profiles, we demonstrate that minor adjustments to the crawling technology (e.g., browser type or mimicking user interaction) may lead to significant differences in the results. We show that inadvertently documented experiments reduce the chances that researchers can reliably reproduce the results. More importantly, *replicability* and comparability of individual works cannot be universally assumed. A takeaway from this is that we as a community need to find ways to perform more robust Web measurements to draw reproducible and replicable conclusions from the conducted experiments. We hope that our results improve the situation and spark a vivid discussion.

Acknowledgments. This work was supported by the Federal Ministery for Economic Affairs and Climate Action (grant 01MK20008E "Service-Meister" and 68GX21006G "TELLUS"), by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2092 *CASA* (390781972), and by the Helmholtz Association (HGF) within topic "46.23 Engineering Secure Systems".

References

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *ACM Conference on Computer and Communications Security*, CCS, 2014. doi:10.1145/2660267.2660347.

[2] Shubham Agarwal and Ben Stock. Critical errors in our recent MADweb paper, 2021. URL: https://swag.cispa.saarland/default/2021/07/19/madweb-headers.html.

[3] Shubham Agarwal and Ben Stock. First, Do No Harm: Studying the manipulation of security headers in browser extensions. In *Workshop on Measurements, Attacks, and Defenses for the Web*, MADWeb, 2021. doi:https://dx.doi.org/10.14722/madweb.2021.23016.

[4] Syed Suleman Ahmad, Muhammad Daniyal Dar, Muhammad Fareed Zaffar, Narseo Vallina-Rodriguez, and Rishab Nithyanand. Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web. In *International Conference on World Wide Web*, TheWebConf, 2020. doi:10.1145/3366423.3380113.

[5] Alexa Internet, Inc. The Top 500 Sites on the Web, 2021. URL: https://www.alexa.com/topsites/.

[6] Mark Alllman and Vern Paxson. Issues and Etiquette Concerning Use of Shared Measurement Data. In *ACM SIGCOMM Internet Measurement Conference*, IMC, 2007. doi:10.1145/1298306.1298327.

[7] Waqar Aqeel, Balakrishnan Chandrasekaran, Anja Feldmann, and Bruce M. Maggs. On Landing and Internal Web Pages: The Strange Case of Jekyll and Hyde in Web Performance Measurement. In *ACM SIGCOMM Internet Measurement Conference*, IMC, 2020. doi:10.1145/3419394.3423626.

[8] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and Don'ts of Machine Learning in Computer Security. In *USENIX Security Symposium*, Usenix Sec., 2022.

[9] Association for Computing Machinery. Artifact Review and Badging Version 1.1, 2020. URL: https://www.acm.org/publications/policies/artifact-review-and-badging-current.

[10] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The Menlo Report. *IEEE Security & Privacy*, 10(02), 2012. doi:10.1109/MSP.2012.52.

[11] Sarah Bird, Ilana Segall, and Martin Lopatka. Replication: Why We Still Can't Browse in Peace: On the Uniqueness and Reidentifiability of Web Browsing Histories. In *Symposium on Usable Privacy and Security*, SOUPS, 2020.

[12] Sam Burnett and Nick Feamster. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In *ACM Conference on Special Interest Group on Data Communication*, SIGCOMM, 2015. doi:10.1145/2785956.2787485.

[13] Stefano Calzavara, Tobias Urban, Dennis Tatang, Marius Steffens, and Ben Stock. Reining in the Web's Inconsistencies with Site Policy. In *Symposium on Network and Distributed System Security*, NDSS, 2021. doi:10.14722/ndss.2021.23091.

[14] 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*, 2(1), 2022.

[15] cURL Development Team. cURL – Command Line Tool and Library for Transferring data with URLs, 2021. URL: https://curl.se/.

[16] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. Measuring Cookies and Web Privacy in a Post-GDPR World. In *Conference on Passive and Active Measurement*, PAM, 2019. doi:10.1007/978-3-030-15986-3_17.

[17] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *ACM Transactions on Information Systems*, 2(39), 2021.

[18] Nurullah Demir, Tobias Urban, Kevin Wittek, and Norbert Pohlmann. Our (in)Secure Web: Understanding Update Behavior of Websites and Its Impact on Security. In *Conference on Passive and Active Measurement*, PAM, 2021. doi:10.1007/978-3-030-72582-2_5.

[19] Clemens Deußer, Steffen Passmann, and Thorsten Strufe. Browsing Unicity: On the Limits of Anonymizing Web Tracking Data. In *IEEE Symposium on Security and Privacy*, S&P, 2020. doi:10.1109/SP40000.2020.00018.

[20] EasyList. EasyPrivacy. https://easylist.to/easylist/easylist.txt, 2021.

[21] Steven Englehardt and Arvind Narayanan. Online Tracking: A 1-Million-Site Measurement and Analysis. In *ACM Conference on Computer and Communications Security*, CCS, 2016. doi:10.1145/2976749.2978313.

[22] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Tracking Personal Identifiers Across the Web. In *pam*, PAM, 2016. doi:10.1007/978-3-319-30505-9_3.

[23] Google Inc. BigQuery: Cloud Data Warehouse. https://cloud.google.com/bigquery, 2021.

[24] Google, Inc. Chromium, 2021. URL: https://www.chromium.org/Home.

[25] Google, Inc. Fix lazy-loaded content, 2021. URL: https://developers.google.com/search/docs/guides/lazy-loading?hl=en.

[26] Jack Hirshleifer. Privacy: Its origin, function, and future. *The Journal of Legal Studies*, 9(4), 1980.

[27] HTTP Archive. The HTTP Archive Tracks How the Web is Built. https://httparchive.org, 2021.

[28] Xuehui Hu, Guillermo Suarez de Tangil, and Nishanth Sastry. Multi-country Study of Third Party Trackers from Real Browser Histories. In *IEEE European Symposium on Security and Privacy*, EuroS&P, 2020. doi:10.1109/EuroSP48549.2020.00013.

[29] Imperva, Inc. Bad Bot Report 2020: Bad Bots Strike Back, 2020. URL: https://www.imperva.com/blog/bad-bot-report-2020-bad-bots-strike-back/.

[30] Luca Invernizzi, Kurt Thomas, Alexandros Kapravelos, Oxana Comanescu, Jean-Michel Picod, and Elie Bursztein. Cloak of Visibility: Detecting When Machines Browse a Different Web. In *IEEE Symposium on Security and Privacy*, S&P, 2016. doi:10.1109/SP.2016.50.

[31] John Ioannidis. Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8), 2005. doi:10.1371/journal.pmed.0020124.

[32] Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian, and Zubair Shafiq. AdGraph: A Graph-Based Approach to Ad and Tracker Blocking. In *IEEE Symposium on Security and Privacy*, S&P, 2020. doi:10.1109/SP40000.2020.00005.

[33] Hugo Jonker, Benjamin Krumnow, and Gabry Vlot. Fingerprint Surface-Based Detection of Web Bot Detectors. In *European Symposium on Research in Computer Security*, ESORICS, 2019. doi:10.1007/978-3-030-29962-0_28.

[34] Jordan Jueckstock, Shaown Sarker, Peter Snyder, Aidan Beggs, Panagiotis Papadopoulos, Matteo Varvello, Ben Livshits, and Alexandros Kapravelos. Towards Realistic and Reproducible Web Crawl Measurements. In *International Conference on World Wide Web*, TheWebConf, 2021. doi:10.1145/3442381.3450050.

[35] Will Keeling. selenium-wire 4.3.1. https://pypi.org/project/selenium-wire/, 2021.

[36] George Klees, Andrew Ruef, Benji Cooper, Shiyi Wei, and Michael Hicks. Evaluating Fuzz Testing. In *ACM Conference on Computer and Communications Security*, CCS, 2018. doi:10.1145/3243734.3243804.

[37] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. Browser Fingerprinting: A Survey. *ACM Transactions on the Web*, 14(2), 2020. doi:10.1145/3386040.

[38] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Symposium on Network and Distributed System Security*, NDSS, 2019. doi:10.14722/ndss.2019.23386.

[39] Bohan Li, Yongxiang Cai, Shuying Deng, and Zongyi He. The Strategy of Personal Customization and Method of Collecting Professional Dynamic Information. In *Journal of Physics: Conference Series*, JPCS, 2020. doi:10.1088/1742-6596/1626/1/012034.

[40] Majestic. The Majestic Million – The million domains we find with the most referring subnets, 2022. URL: https://majestic.com/reports/majestic-million/.

[41] Sourena Maroofi, Maciej Korczyński, and Andrzej Duda. Are You Human? Resilience of Phishing Detection to Evasion Techniques Based on Human Verification. In *ACM SIGCOMM Internet Measurement Conference*, IMC, 2020. doi:10.1145/3419394.3423632.

[42] Célestin Matte, Nataliia Bielova, and Cristiana Santos. Do Cookie Banners Respect my Choice? : Measuring Legal Compliance of Banners from IAB Europen's Transparency and Consent Framework. In *IEEE Symposium on Security and Privacy*, S&P, 2020. doi:10.1109/SP40000.2020.00076.

[43] Mozilla Foundation,. OpenWPM on GitHub, 2021. URL: https://github.com/mozilla/OpenWPM.

[44] National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC, 2019. URL: https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science, doi:10.17226/25303.

[45] Arian Akhavan Niaki, Shinyoung Cho, Zachary Weinberg, Nguyen Phong Hoang, Abbas Razaghpanah, Nicolas Christin, and Phillipa Gill. ICLab: A Global, Longitudinal Internet Censorship Measurement Platform. In *IEEE Symposium on Security and Privacy*, S&P, 2020. doi:10.1109/SP40000.2020.00014.

[46] Łukasz Olejnik, Claude Castelluccia, and Artur Janc. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. In *Proceedings on Privacy Enhancing Technologies*, PETS, 2012.

[47] Andriy Panchenko, Fabian Lanze, Jan Pennekamp, Thomas Engel, Andreas Zinnen, Martin Henze, and Klaus Wehrle. Website Fingerprinting at Internet Scale. In *Symposium on Network and Distributed System Security*, NDSS, 2016. doi:10.14722/ndss.2016.23477.

[48] Vern Paxson. Strategies for Sound Internet Measurement. In *ACM SIGCOMM Internet Measurement Conference*, IMC, 2004. doi:10.1145/1028788.1028824.

[49] PostgreSQL Global Development Group. PostgreSQL: The World's Most Advanced Open Source Relational Database. https://www.postgresql.org/, 2021.

[50] James H Price and Judy Murnan. Research Limitations and the Necessity of Reporting them. *American Journal of Health Education*, 35(2), 2004.

[51] Proton Technologies AG. ProtonVPN: Secure and Free VPN service for protecting your privacy, 2021. URL: https://protonvpn.com/.

[52] Paula T Ross and Nikki L Bibler Zaidi. Limited by our Limitations. *Perspectives on Medical Education*, 8(4), 2019.

[53] Christian Rossow, Christian J. Dietrich, Grier Grier, Christian Kreibich, Vern Paxson, Norbert Pohlmann, Herbert Bos, and Maarten van Steen. Prudent Practices for Designing Malware Experiments: Status Quo and Outlook. In *IEEE Symposium on Security and Privacy*, 2012. doi:10.1109/SP.2012.14.

[54] Patrick E. Shrout and Joseph L. Rodgers. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, 69(1), 2018. doi:10.1146/annurev-psych-122216-011845.

[55] Software Freedom Conservancy. SeleniumHQ Browser Automation. https://www.selenium.dev/, 2021.

[56] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Beyond the Front Page: Measuring Third Party Dynamics in the Field. In *International Conference on World Wide Web*, TheWebConf, 2020. doi:10.1145/3366423.3380203.

[57] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Measuring the Impact of the GDPR on Data Sharing. In *ACM Asia Conference on Computer and Communications Security*, AsiaCCS, 2020. doi:10.1145/3320269.3372194.

[58] Erik van der Kouwe, Gernot Heiser, Dennis Andriesse, Herbert Bos, and Cristiano Giuffrida. SoK: Benchmarking Flaws in Systems Security. In *IEEE European Symposium on Security and Privacy*, EuroS&P, 2019. doi:10.1109/EuroSP.2019.00031.

[59] Benjamin VanderSloot, Allison McDonald, Will Scott, J. Alex Halderman, and Roya Ensafi Ensafi. Quack: Scalable Remote Measurement of Application-Layer Censorship. In *USENIX Security Symposium*, Usenix Sec., 2018.

[60] Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Xavier Blanc. FP-Crawlers: Studying the Resilience of Browser Fingerprinting to Block Crawlers. In *Workshop on Measurements, Attacks, and Defenses for the Web*, MADWeb, 2020. doi:https://dx.doi.org/10.14722/madweb.2020.23010.

[61] John-Paul Verkamp and Minaxi Gupta. Inferring Mechanics of Web Censorship Around the World. In *Workshop on Free and Open Communications on the Internet*, FOCI, 2012.

[62] David Y. Wang, Stefan Savage, and Geoffrey M. Voelker. Cloak and Dagger: Dynamics of Web Search Cloaking. In *ACM Conference on Computer and Communications Security*, CCS, 2011. doi:10.1145/2046707.2046763.

[63] Peter Wurzinger, Christian Platzer, Christian Ludl, Engin Kirda, and Christopher Kruegel. SWAP: Mitigating XSS Attacks Using a Reverse Proxy. In *ICSE Workshop on Software Engineering for Secure Systems*, IWSESS, 2009. doi:10.1109/IWSESS.2009.5068456.

[64] Zhiju Yang and Chuan Yue. A Comparative Measurement Study of Web Tracking on Mobile and Desktop Environments. In *Proceedings on Privacy Enhancing Technologies*, PETS, 2020. doi:https://doi.org/10.2478/popets-2020-0016.

[65] David Zeber, Sarah Bird, Camila Oliveira, Walter Rudametkin, Ilana Segall, Fredrik Wollsén, and Martin Lopatka. The Representativeness of Automated Web Crawls as a Surrogate for Human Browsing. In *International Conference on World Wide Web*, TheWebConf, 2020. doi:10.1145/3366423.3380104.

A Availability of Data & Code Artifacts

To foster future research, we release our code, measurement data, and other supplementary information openly online at: https://github.com/awareseven/Reproducibility-and-Replicability-of-Web-Measurement-Studies

B Individual Criterion Analysis

The results of the analysis for each of our 18 criteria are given in Table 2. Roughly 40% of the criteria have been satisfied, 7% were not properly documented, 50% were omitted, and 2% were not applicable to a paper. The decision if a criterion is (fully) satisfied is not always binary. Especially the distinction between "N/A" and "omit" is sometimes not straightforward. For example, for C12, we rated LePochat et al.'s work, which builds the Tranco list [38], with "N/A". We have done so because the geolocation has no impact on the work (i.e., it

does not matter from which location one reads the toplist). An example for a more complex process is C13, where we rated the work of Matte et al., where they analyze whether the choice of cookie banners is respected [42]. We decided on "N/A" because the study focuses on the respect of choice, and not whether or how the respect is treated in different contexts.

C Experimental Setup

To measure the impact of different experimental setups, we built a pipeline that enables us to compare measurement results based on different setups. To allow comparability, parallel page visits across all defined profiles are essential. Our measurement setup consists of different virtual machines (VMs) orchestrated by a "commander" to organize parallel page visits. For all VMs, we use *Ubuntu 20.04* as operating system and do not pass the GPU to them, which can impact fingerprinting scripts [1]. Each of the worker VMs conducts the measurement according to a specific profile (see C5–C10 below).

As an initial step, the commander assembles the set of URLs that should be visited during the experiment based on the heuristic described in C1 and C2 (see below). For our experiment, this had happened on 06/24/2021, three days before we started the measurement. At the beginning of the measurement, the commander starts one VM for each of the 24 browser profiles. The VMs will query the commander for batches of sites ($n = 10$) to analyze. All VMs receive an identical list of sites and pages, such that all VMS conduct measurements in the same order, starting with the site's landing page. Once all managed VMs are ready to start their measurements, the commander issues a signal to start the experiment as *stateless* coordinated crawls of the provided URLs. Each VM starts 10 browsers (one for each site) in parallel using the defined profile. Once the analysis of a page is finished, the same browser instance is moving to the next page of the site. Once the results have been stored, the VM will query for the next batch of URLs and wait until the commander tells all VMs to start the analysis. To conduct this large-scale measurement and to host the virtual machines with the necessary resources, we use a server which is equipped with 256GB RAM, a *AMD 7542, 2.90GHz* CPU with 32 cores, and a 10 Gbps network interface. We supplied each VM with 10GB RAM, five CPUs, and 40GB of hard disc space to cache the results before sending them to the commander.

C.1 Dataset

C1 In our analysis, we use the Tranco list generated on 06/23/2021, which is available at `https://tranco-list.eu/list/ZGPG/15000` [38]. We use the top 10k sites to build our website dataset.
C2 Tranco lists only contain sites (eTLD+1). Therefore, we used the following heuristic to identify the landing page by defining a protocol to use. We used four prefixes (`http[s]://[www.]`, starting with the https variants) and test if the resulting URL is reachable. If so, we added the first identified site to our corpus. Once we determined the seed URLs for our crawl, we visited each of them and randomly chose up to 25 first-party links (recursively if necessary), which we used for our measurement run. Selecting subsites is essential since they often show a different behavior compared to the landing pages [56, 7]. To avoid incorporating duplicated URLs in our dataset, we always ignore each identified link's anchor part and omit links that could result in redirects (e.g., links that contain `http://` in the path).
C3 We make the list of pages and sites openly available (see App. A).
C4 For our continuous measurement, we visit the top 1k sites from our website corpus (18,377 distinct pages) daily (starting at midnight) throughout our experiment (from 07/08/2021 to 07/19/2021).

C.2 Experimental Design

C5 We use four different browser types (Chrome, Chrome headless, Firefox, Firefox headless) in our study. We use the popular *OpenWPM* Framework [21] (v0.15.0 – Firefox version 88) to perform the Firefox-based measurements and Chromium [24] to perform the Chrome-based measurements, respectively.
C6 Regarding adjustments to *OpenWPM*, we built a wrapper that feeds the pages to visit and which extracts the measurement results. Hence, the wrapper does not affect the functionality of the framework. Regarding changes to *OpenWPM*, we use two custom commands provided by the framework (1)

for logging visits and (2) to simulate "user interaction" (see C10). Regarding flexible options of the framework, we used the `native` display mode, which enables the GUI browser and disable it for the the `headless` mode. Otherwise, we used the standard configuration of the platform. For our Chrome setup, we aim to use setups similar to other studies [4]. Hence, we utilize *Selenium* [55] to build our Chrome-based crawler (version: 91). When implementing the crawler and page visiting strategy, we oriented at the parameters used in *OpenWPM* to allow more realistic comparison (e.g., each page visit is done in a new tab or waiting for resources to be loaded). To conduct the headless crawl, we pass the *headless* argument to Selenium, which causes Chrome to start in the headless mode.

C7 Aside the mentioned adjustments, we did not extend the browsers.

C8 To disguise our crawler, we modified the standard *Selenium* parameters based on the findings of Jonker et al., who empirically studied which techniques are used in practice to detect such crawlers [33] (e.g., changing the user agent). Otherwise, we implemented simulated user interaction (see C10).

C9 We make the used framework publicly available (see App. A).

C10 We use two approaches to simulate user interaction: (1) no interaction ("none") and (2) mimicking artificial user interaction ("user interaction"). Thus, for (1) we do not interact with the website at all, besides from waiting for the site to finish loading. In profile (2), once the browser loads the page, we wait for 30 seconds or until the page finished loading and then simulate three *page down* keystrokes followed by three *Tab* keystrokes, and finally an *end* keystroke with minimal periods of delay in between.

C11 When visiting a page, we wait until a page has finished loading, close the browser, and move on to the next page. We use a timeout of 30 seconds for each page visit (standard configuration of *OpenWPM*), after which the visit will be terminated (e.g., to avoid slow websites or other problems that will delay the measurement process).

C12 We choose three geolocations for our measurement: (1) Germany (EU), (2) Japan (AS), and (3) the United States (NA). We choose these continents since the majority of the analyzed papers focuses on them and we choose Japan to avoid bias due to censorship [61]. To simulate different geolocations, we used *ProtonVPN* [51].

C13 We altered browsers resolution to `1366x768` and the user agents to `Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/ 91.0.4472.77 Safari/537.36` for *Chrome* and `Mozilla/5.0 (X11; Linux x86_64; rv:88.0) Gecko/20100101 Firefox/88.0` for *Firefox*.

C14 For *OpenWPM*, after crawling a site, our wrapper (described in C6) extracts the HTTP traffic from the database provided by *OpenWPM*. The Chrome-based crawlers collect the HTTP traffic with the help of *selenium-wire* [35]. The results of the measurements are pushed into a *Google BigQuery* database [23] and meta information is stored in a local *PostgreSQL* database [49].

C.3 Evaluation

C15 Our (raw) results are publicly available (see App. A).

C16 We elaborate on our findings in Sec. 4.2.

C17 & C18 We discuss limitations and ethics in Sec. 6.

Table 2: Overview of our survey's results.

	ID	N/A	Omit	Undoc.	Sat.	Interpretation
Dataset	C1	0%	3%	32%	64%	A third of the analyzed papers did not document which sites they analyzed, which makes an reproduction of the results nearly impossible.
	C2	0%	73%	5%	22%	The vast majority (73%) of works omit or do not document which pages they analyzed. However, not all pages on a same site show the same behavior [56, 13].
	C3	0%	32%	13%	56%	The documentation which pages are analyzed is somewhat balanced. The works that did not document or omitted this step (62%) presumably analyzed the landing page of a site.
	C4	1%	62%	3%	33%	Nearly two-thirds of the papers use a single snapshot to analyze a phenomenon. If research checks for a certain behavior (e.g., a vulnerability), temporal trends should be included.
Experimental Design	C5	2%	31%	10%	57%	Approx. two-fifth of the papers completely omit or do not document the crawler properly, which is a problem because the used crawled has an direct impact on the produces results [4].
	C6	3%	50%	8%	40%	The majority of the papers (60%) of the papers do *not* state, which adjustments they made to the crawler configuration. This has a negative impact on reproducibility of the results since one cannot rebuild the described setup.
	C7	6%	60%	5%	29%	Only Two-fifth of the analyzed papers do state whether they developed an extension or if they extended the crawler at all. The impact on the reproducibility is similar to C6.
	C8	3%	88%	0%	9%	The majority of papers do not state which approaches (if any) were taken to evade bot detection. Not taking any action to avoid detection can impact the results significantly [56, 34].
	C9	6%	67%	1%	29%	Almost a third of the analyzed papers make their crawler available. This is making reproducibility harder, does not allow comparison to previous work, and also contradicts open science.
	C10	4%	67%	4%	25%	Only roughly a third of the papers took steps how they mimicked user interaction. Similarly to the used crawling strategy (C11) user interaction verifiable impact the outcomes of results [56].
	C11	3%	41%	12%	44%	More than half of the analyzed papers provided details on their crawling strategy. However, the vast majority does not provide details on the process or omits the description. The crawling strategy might have a significant impact on the results(e.g., visited pages [7]).
	C12	2%	72%	2%	25%	Most papers do not state from which location the measurement was performed. The geographical location of a measurement is important information since the results might be impacted by for example censorship [59] or different legislation [57].
	C13	3%	70%	4%	23%	A majority of papers do not state if and which adjustments they made to the used browser. Depending on the research question already small adjustments can highly impact the results (e.g., using a headless browser to scale up the experiment [34]).
	C14	0%	14%	10%	76%	A majority of papers describes their data processing pipeline, which enables the reproducibility or replicability of the experiment. However, 36% of the papers lack that type of information.
Evaluation	C15	0%	74%	3%	24%	More than two out of three papers do not provide raw measurement results. Not providing these results makes the reproducibility of the results harder.
	C16	0%	1%	1%	98%	Nearly all papers describe the outcome of the measurement on a higher level and provide general results. This is an expectable result.
	C17	0%	35%	2%	63%	Most papers provide a limitations section, but more than one-third do not discuss limitations.Providing details on limitations is not only good scientific practice but also important to assess the results of a study.
	C18	0%	64%	1%	35%	Almost two-thirds of the measurement papers do not provide an ethics section. Since they focus on security & privacy issues the ethics of each experiment should be discussed [10].