Do Opt-Outs Really Opt Me Out?

Duc Bui
University of Michigan
Ann Arbor, USA
ducbui@umich.edu

Brian Tang
University of Michigan
Ann Arbor, USA
bjaytang@umich.edu

Kang G. Shin
University of Michigan
Ann Arbor, USA
kgshin@umich.edu

ABSTRACT

Online trackers, such as advertising and analytics services, have provided users with choices to opt out of their tracking and data collection to mitigate the users' concerns about increased privacy risks. While opt-out choices of online services for the cookies placed on their own websites have been examined before, the choices provided by trackers for their third-party tracking services on publisher websites have been largely overlooked. There is no guarantee that a tracker's opt-out options would faithfully follow the statements in its privacy policy. To address this concern, we develop an automated framework, called `OptOutCheck`, that analyzes (in)consistencies between trackers' data practices and the opt-out choice statements in their privacy policies. We create sentence-level classifiers, which achieve ≥84.6% precision on previously-unseen statements, to extract the opt-out policies that state neither tracking nor data collection for opted-out users from trackers' privacy-policy documents. `OptOutCheck` analyzes both tracker and publisher websites to detect opt-out buttons, perform the opt-out, and extract the data flows to the tracker servers after the user opts out. Finally, we formalize the opt-out policies and data flows to derive logical conditions to detect the inconsistencies. In a large-scale study of 2.9k popular trackers, `OptOutCheck` detected opt-out choices on 165 trackers and found 11 trackers who exhibited data practices inconsistent with their stated opt-out policies. Since inconsistencies are violations of the trackers' privacy policies and demonstrate data collection without user consent, they are likely to lose users' trust in the online trackers and trigger the necessity of an automatic auditing process.

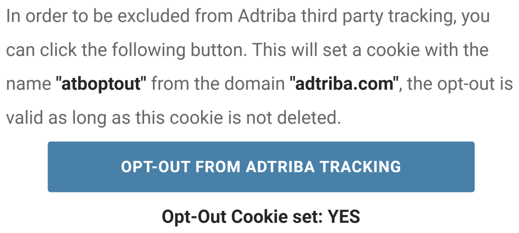
CCS CONCEPTS

• **Security and privacy** → **Usability in security and privacy**; **Privacy protections**.

KEYWORDS

online tracker; opt-out choice; privacy policy; consistency analysis

ACM Reference Format:
Duc Bui, Brian Tang, and Kang G. Shin. 2022. Do Opt-Outs Really Opt Me Out?. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22), November 7–11, 2022, Los Angeles, CA, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3548606.3560574



In order to be excluded from Adtriba third party tracking, you can click the following button. This will set a cookie with the name **"atboptout"** from the domain **"adtriba.com"**, the opt-out is valid as long as this cookie is not deleted.

Opt-Out Cookie set: YES

Figure 1: Example opt-out setting and policy statements. A user opts out of tracking by clicking the opt-out button that creates a cookie to record the user's opt-out choice.

1 INTRODUCTION

Online trackers, such as ad platforms and analytics service providers, leverage various tracking techniques to collect users' browsing history across websites, posing serious privacy concerns to users and regulators. As a result, the trackers' privacy policies often provide users with an opt-out link or button to reject targeted advertisements and/or their data collection [28, 66]. Fig. 1 shows an example where an ad platform states to stop tracking users via unique-identifier cookies after the user opts out.

Inconsistencies between a stated opt-out policy and its actual tracking behavior pose high privacy risks to users since the data collection occurs/continues even after they opt out, contrary to their expectations. These inconsistent privacy practices can also be deemed deceptive and illegal by regulators. The Federal Trade Commission (FTC) has fined several ad networks for their short-lived opt-out cookies [23], deceptive policy statements about a complete cookie opt-out [25], and falsified statements on browser cookie settings [24]. Therefore, checking (in)consistencies between the stated privacy policies and the corresponding data practices is important as it benefits all of the users, companies and regulators; users will be reassured of their privacy protection, regulators can prevent trackers' deceptive mechanisms, and tracker companies will be forced to comply with their stated privacy policies.

The main research question to answer is then: *Do opt-out settings really opt users out of an online tracker's data practices as stated in its opt-out policy?* To answer this question, we address the following three challenges that originate from the complexity and vagueness of the opt-out policies specified in legal language and the variability of non-standardized opt-out links/buttons. First, the semantic extraction and analysis of opt-out policy statements are difficult due to the complexity of the website user interface and the legal language used in privacy policies. Second, analyzing the data collection and tracking behavior requires activating an opt-out choice, extracting data flows and inferring data-usage purposes of trackers after the opt-out setting is enabled. Finally, verifying (in)consistencies between the opt-out policies and the data-collection practices needs to reconcile the different (i.e., high vs. low) levels of granularity between the policy statements and data flows.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CCS '22, November 7–11, 2022, Los Angeles, CA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9450-5/22/11…$15.00
https://doi.org/10.1145/3548606.3560574

Unlike prior studies on the opt-out choices provided by content-publishing websites, we study trackers' opt-out of tracking services as third parties on the content websites. Prior work [12, 50, 51, 85] has mainly studied the usability of opt-out choices and the extraction of generic opt-out hyperlinks on content-providing websites, rather than direct opt-out settings of online trackers. A recent study of the compliance of cookie banners [76] does not apply to the cookies on websites *other* than those hosting the banners, thus covering a different scope from our work. Moreover, none of prior studies has checked the (in)consistencies between the opt-out settings and privacy policies. They assumed that trackers always honored users' opt-out preferences once the opt-out cookies were set [28, 66].

To fill these gaps, we present OptOutCheck, an automated framework that analyzes (in)consistencies between opt-out policy statements and the corresponding data practices of online trackers.

First, given a tracker's website, OptOutCheck extracts its opt-out buttons that record a user's preference of opting out of the tracker's tracking and data collection. The system extracts policy statements about the privacy practices for opted-out users (called *opt-out policies*). It identifies 5 opt-out policy classes (e.g., *No-tracking* and *No-data-collection*) by analyzing semantic arguments, syntactic dependencies and text patterns of the policy sentences. For example, a tracker may not use unique-ID cookies to track opted-out users.

Second, OptOutCheck extracts the data flows from a user's browser to a tracker's servers after the user activates the opt-out choices. To this end, OptOutCheck simulates a user's click on opt-out buttons, identifies opt-out cookies and determines the cookie domains enforced by the opt-out policies. OptOutCheck then identifies the tracking and data-collection behavior by analyzing the data types and usage purposes of the key–values sent via cookies and URL parameters to the tracker's servers after an opt-out.

Finally, we formalize policy statements, data flows and subsumptive relationships of data types to define the condition under which a data flow is consistent with a privacy policy. OptOutCheck checks this condition based on opt-out policies and data types of flows to detect flow-to-policy inconsistencies. Inspired by the soundness of dynamic analysis in software testing [53, 54, 98], we aim to minimize false positives so that the reported inconsistencies should always be true positives. In a large-scale study, OptOutCheck found multiple inconsistencies of popular online trackers which we manually verified, demonstrating OptOutCheck's scalability and effectiveness.

This paper makes the following main contributions:

- Classification of opt-out policies and creation of automatic classifiers for policy sentences. We categorize policy statements — which describe the data-collection policies after a user opts out — into 5 policy classes. We create a dataset (available at [16]) and classifiers based on natural language processing (NLP) that achieve a ≥84.6% precision on previously-unseen samples.
- Extraction of data-collection behavior of trackers after a user opts out. We create a dataset and derive a classifier to identify opt-out cookies that achieves 95% precision on the test set. We develop techniques to extract the scope of opt-out policies based on opt-out cookies, extract the matching data traffic and infer the data types collected by a tracker.
- A formal analysis of (in)consistencies between the opt-out policy statements and data flows conditioned on users' opt-out (Section 8). We derive formal consistency conditions and logical rules to detect the inconsistencies based on the classification of opt-out policies and data flows.
- An end-to-end (E2E) automated framework, OptOutCheck, that detects (in)consistencies between the actual data practices and the stated opt-out policies of online trackers.
- A large-scale study of opt-out choices of 2,981 online trackers. Of the 165 trackers for which OptOutCheck detected opt-out buttons and opt-out cookies, 11 trackers were found to track and collect user data despite their policy statements to stop the tracking and/or data collection after the user's opt-out. These trackers were present on 3.65% of the top 10k websites on average and tracked a significant amount of web traffic. Since the inconsistencies are direct violations of the trackers' own privacy policies while the trackers collected user data without the users' consent, regulators may impose heavy fines for their deceptive privacy practice and unlawful data collection.

The rest of the paper details OptOutCheck's analysis pipeline (Fig. 2). OptOutCheck first searches for opt-out buttons on privacy-policy web pages (Section 4). It then extracts the corresponding opt-out policy statements (Section 5), opt-out cookies (Section 6), and data flows after opting out (Section 7). Finally, the system checks the conditions to detect inconsistencies, if any (Section 8).

2 BACKGROUND

2.1 Trackers and Tracking Mechanisms

Trackers are companies that collect information about users who browse the web [44]. The most common types are advertisers and data analytics services that collect user data to deliver online behavioral advertising (OBA). Similarly, site analytics and social media track users to understand user-activity patterns to improve and provide their services [32, 43]. As depicted in Fig. 3, we consider data flows among users, trackers and publisher websites. When a user accesses a content-providing website, besides the publisher's own contents the user wants to read, the browser also loads trackers' cookies and scripts. Trackers offer users opt-out choices on their websites so that they can request not to track or collect their data.

The most common online tracking technology used in practice and stated in privacy policies is HTTP cookies placed on user devices [66, 91]. Members of the Digital Advertising Alliance (DAA) in the USA and Canada agree not to use Flash and similar local-storage-based tracking tools unless an opt-out mechanism is publicly provided [20]. Other advanced mechanisms are harder to detect, e.g., canvas fingerprinting, ever cookies, and cookie syncing [2].

We consider *third-party cookies* that are the cookies in domains other than those of the websites being accessed regardless of domain ownership [19, 41, 103]. We use the term "domain" to indicate a pay-level domain that a consumer or business can directly register, and is typically a subdomain followed by an effective top-level domain (public suffix) [40, 70] which can be extracted by *tldextract* [67].

2.2 Opt-out Mechanisms

Placing anonymous opt-out cookies in the users' web browsers to signal their choices is the *de facto* mechanism used by trackers [66]. It is possible to have a persistent identifier for opt-out purposes, but trackers can now easily track users who contradict the purpose

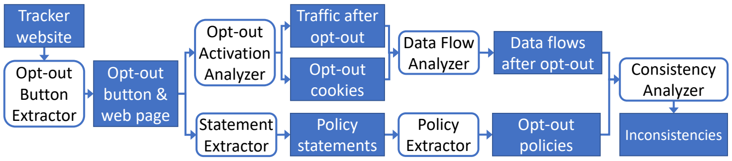


Figure 2: OptOutCheck workflow.

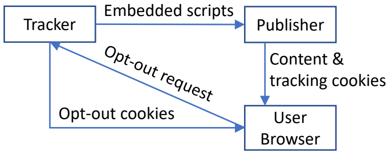


Figure 3: Trackers' data flows.

of opt-out. Many trackers' privacy policies even describe their opt-out mechanisms explicitly [66], such as 'this will set a cookie with the name "atboptout" from the domain "adtriba.com"' as depicted in Fig. 1. Furthermore, tracking blocking tools, such as those of Network Advertising Initiative (NAI) [57], DAA [20], and Evidon Global Opt-out [29]), use this method.

Anonymous opt-out cookies remain the most common opt-out mechanism for advertisers [11, 92] and were explicitly described in the privacy policies we surveyed, and hence we only consider cookie-based tracking and opt-out mechanisms. Although other forms of tracking like fingerprinting exist, fingerprinting is not stable owing to the changes of user fingerprints over time, so trackers even employ cookie re-spawning to enable reliable tracking of users [39]. Another opt-out mechanism uses server-side storage to store user consents [37]. However, it requires a long-term ID for each user, such as the user's ID or email address, and needs to perform synchronization between server-side consent storage and cached local cookies on the user's browser. Since we consider the opt-out settings that do not require a user to log in or input his/her email address, this opt-out mechanism is outside of our scope.

3 COOKIE CRAWLER

We developed a crawler based on Playwright [78] that automates the Google Chrome browser to visit web pages, perform user interaction and record HTTP cookies set by both JavaScript and HTTP responses. To reduce measurement bias due to websites' bot mitigation [63], the crawler utilized an 8-node cluster located in our university and emulated realistic human browsing behavior to circumvent trackers' bot-protection mechanisms [14].

Each web page visit waits until there is no network activity for at least 0.5 seconds or a 30-second timeout expires, which is a common heuristic used by web automation tools for loading dynamic web pages [47, 78]. (See Appendix A for the rationale of the loading timeouts.) Furthermore, if the loading fails, to avoid transient network errors, the web page load was retried at most three times with a 2-minute waiting time between two retries.

4 EXTRACTION OF OPT-OUT BUTTONS

This section describes the detection of the actionable choices provided by trackers for users to opt out of their data collection and tracking. We define an *opt-out activation button* (also called an *opt-out button*) as a clickable HTML element that, upon its click, will record the user's preference of opting out of the trackers' services. Similarly, an opt-out page is a web page that contains an opt-out button. Such pages can be an iframe embedded in another web page. Furthermore, while many websites instruct users to use opt-out tools providing self-regulatory groups such as NAI [58] and DAA [7], OptOutCheck does not analyze them because they do not contain any specific definition of a tracker's opt-out. These groups' members frequently provide their own opt-out definitions which are stricter than the minimum requirements of NAI and DAA [66].

Given a tracker domain, OptOutCheck uses a three-stage pipeline to extract its opt-out button. It first identifies the candidate web pages that may contain an opt-out button or a link to an opt-out page by searching for keywords related to "*opt out*" in the entire website. OptOutCheck then detects opt-out button candidates from the web pages. Finally, OptOutCheck validates opt-out buttons by extracting the opt-out cookies after clicking the candidate buttons.

4.1 Extraction of Opt-out Page Candidates

4.1.1 Challenges. As trackers have incentives to keep users from opting out of their tracking [62], they tend to make it difficult to detect opt-out pages on their websites. The opt-out pages can be placed deep down in the website's hierarchy with very few links to the pages. An advertiser's website may also have multiple policies, such as privacy and cookie policies, but only one of them has a link that points to its opt-out page. Similarly, searching multilingual websites requires discovering their language-switching links.

Because checking the availability of an opt-out page requires exhaustive crawling of the whole website, we leverage search engines that systematically index web pages of trackers' websites to find opt-out page candidates. Although search engines may not crawl all websites in real time, the privacy policies and opt-out links do not change very frequently [8]. Finally, the results in this step are refined further in other detection steps in the OptOutCheck's pipeline, thus avoiding/minimizing potential false positives.

4.1.2 Query Term Design. We derive a query term for Google Programmable Search Engine [49] to search for the web pages that contain keywords related to the opt-out of trackers' websites. Specifically, we use the query term *"opt out opt-out site:<tracker-domain>"* where the *<tracker-domain>* is substituted for the website domain of a tracker, such as *site:adblade.com*. The query includes *opt out* without any quotes to search for variations of *opt out* such as *opted out* or *opting out*. The term "opt-out" helps detect opt-out pages with that term appearing on their URLs instead of their contents. The search engine then looks for these "opt out" variations and the exact "opt-out" term in both the URL and website's content [80, 95].

The query is designed to have better coverage rather than maximizing precision because the later steps in the pipeline (e.g., opt-out button detection) will filter out unrelated non-opt-out pages. So, the query term avoids restricting the search with the *exactTerms* or *orTerms* parameters [48]. We also try to use the minimum number of customized parameters as using more parameters is found to make the output results less stable over time. For example, restricting to English-only web pages produced no results in some query executions. The Google search may still miss web pages, but it will only increase false negatives (i.e., no detection of opt-out buttons) without increasing false positives (i.e., incorrect detection).

4.1.3 Evaluation. We evaluate the extraction performance on trackers that are known to provide opt-out buttons. Specifically, we randomly selected 100 trackers from the Evidon Global Opt-out list [29]. We excluded inaccessible opt-out pages, possibly due to the outdated opt-out-page URLs in the Evidon database. Finally, we extracted opt-out pages of 43 trackers to create the dataset.

We observed that the search engine is effective in finding the opt-out pages. The opt-out pages are included in the top-1 and top-3 results in 34/43 (79.07%) and 40/43 (93.02%) of the search queries, respectively. There are three cases where the search engine could not detect the opt-out pages. A website places its privacy policy in PDF where the opt-out link is not clickable. Another non-English tracker uses "don't track" instead of the "opt out" keyword. Finally, one website disallows crawling of its privacy policy using *robots.txt* specification [55]), thus preventing search engines and automated web crawlers from detecting the opt-out button placed in the privacy policy. Because the results lower than the top 3 did not improve the opt-out page detection, OptOutCheck uses only the top-3 results from the search engine for further analysis.

4.2 Opt-out Button Detection

We derive patterns to extract opt-out button candidates by following the Snowball bootstrapping procedure which has been widely adopted for extracting information in web and mobile environments [3, 52, 60, 61]. Specifically, we construct patterns of the attribute values of the HTML elements that represent opt-out buttons. A key step is that after each iteration, only the most reliable rules are kept for the next iteration. Therefore, the set of extraction rules improves as it iterates. This is detailed in Algorithm 1.

Let E be a set of extraction rules where each rule $e \in E$ is a tuple of *(element-selector, attribute, value pattern)* that matches the *value pattern* with the value of an *attribute* of the elements selected by the CSS *element-selector*. An *attribute* is an HTML tag's attribute or text content. To avoid mixed effects on different types of HTML elements, each of the rules applies to one tag and one attribute. Specifically, an *element-selector* is a CSS selector that selects only one type of HTML element rather than a list that selects different HTML tags. Similarly, an *attribute* denotes a single attribute of an HTML element rather than an attribute list. The *value pattern* can be a regular expression or a function that performs complex matching on the element's attribute value. An example rule is *('a', text-content, '^opt[-]out')* that matches any anchor element (i.e., hyperlinks) with text content starting with either "opt-out" or "opt out". The regular expression matching is case-insensitive to handle varied capitalization in opt-out buttons' labels. An element's text content represents only human-readable text, not invisible elements [83].

The seed set contains 4 rules to extract elements *a*, *button*, *input* and *span* with text content starting with "opt out" or "opt-out". These HTML tags are commonly used to implement buttons in web pages [22, 73]. As in prior research [60], we also observe that the seed rule set does not significantly affect the final rules if the matching frequency thresholds are tuned properly.

Following the bootstrapping algorithm, we added patterns that use the *id*, *class*, *value*, *onclick* and *href* attributes of these elements. The final extraction rule set contains 14 rules with a frequency cutoff threshold of 10 (i.e., the rules with <10 matches are excluded). We use two patterns: starting with "opt out" variants and contain the "opt out" function identifier (e.g., "optoutToggle"). The matching patterns use dashes, underscores and spaces as the delimiters.

Algorithm 1 OptOutCheck's bootstrapping procedure for extracting opt-out buttons from a large corpus of web pages.

```
Initialize E to a set of seed extraction rules
While E does not grow
    Use the rule set E to detect opt-out buttons
    Generate new rules based on the detected buttons
    Keep only reliable rules; the resulting rule set is E'
    Set E = E'
Output: a set of rules to extract opt-out buttons
```

4.3 Opt-out Choice Activation

To activate an opt-out choice, OptOutCheck attempts to click the opt-out button candidates until an opt-out cookie is detected or a maximum of 5 candidates have been tried. If clicking a link does not create an opt-out cookie, the crawler returns to the original page and tries the next candidate button. Appendix B describes the implementation of a button-clicking action while Section 6 introduces the definition of opt-out cookies.

To reduce the number of link-clicks, OptOutCheck ranks the opt-out button candidates based on the classifier's confidence (i.e., the classification probability). The system prioritizes the matched patterns on the displayed text content which is a user-facing feature. Furthermore, OptOutCheck excludes the candidates based on the URLs that are informational opt-out web pages commonly used by trackers or industrial opt-out tools, such as the DAA and NAI websites. Similarly, hyperlinks that point to the currently visiting page are also removed.

5 OPT-OUT POLICY ANALYSIS

This section describes the automated extraction of opt-out policies from the opt-out web pages of trackers. Automated analysis of opt-out policies is necessary because manual inspection is impractical to cover thousands of advertisers' privacy policies and account for their regular/frequent updates.

5.1 Interpretation and Formal Definitions

5.1.1 Interpretation of Opt-out Policies. We consider an opt-out statement to be equivalent to a negative-sentiment statement, i.e., a statement "opt out of S" is equivalent to "not S after opt-out" where S is a statement about data collection. For example, "you can opt out of receiving targeted ads" is equivalent to "you will not receive targeted ads after opt-out."

Due to the ambiguity of the language used in privacy policies, we make the following interpretation of opt-out statements. Like the interpretation in prior work [66], we assume "no tracking" to indicate that user data can still be collected but will not be associated with the device, such as by using unique-ID cookies. Tracking can be defined as "collecting data over multiple different web pages and sites, which can be linked to individual users via a unique user identifier" [65]. Moreover, we interpret the "targeting" term as "targeted advertising," so "opting out of targeting" means that interest-based advertising will not be displayed to the users.

Opt-out Policy Class	Policy Statement Set
No-tracking	{ ((r, collect, d), (d, not_for, tracking) }
No-data-collection	{ ((r, not_collect, d), None) }
No-data-collection-for-oba	{ ((r, collect, d), (d, not_for, targeted_ad)) }

Table 1: Opt-out policy classes and the corresponding sets of policy statements. The data type *id_data* $\equiv_\delta$ "unique identifier", *d* $\equiv_\delta$ "data", and receiver *r* $\equiv_\delta$ "first party" under an ontology δ. "oba" stands for online-behavioral advertising.

Since a cookie is always sent to its ad provider's server whenever the browser makes a request to the server [13], if the advertiser states that it will stop placing cookies on the user's browser (except for the opt-out preference cookies), or the user can opt out of the advertiser's cookies, we interpret it as equivalent to "stop data collection via *third-party cookies*".

5.1.2 Formal Definitions. Inspired by prior work [9, 18], we formalize statements in privacy policies to analyze the (in)consistencies between the policies and actual data-collection behavior as follows.

Definition 5.1 (Policy Statement)**.** *A policy statement is a pair (dc, du) where dc represents data collection and du is data usage. $dc = (r, c, d)$ denotes whether a receiver r does or does not collect ($c \in$ {collect, not_collect}) a data object d. $du = (d, k, p)$ represents whether a data object d is used for or not for ($k \in$ {for, not_for}) a data usage purpose p of the receiver.*

Definition 5.2 (Semantic Equivalence)**.** *x and y are semantically equivalent, denoted as $x \equiv_o y$, if and only if they are synonyms defined under an ontology o. Similarly, $x \not\equiv_o y$ denotes nonequivalent concepts in an ontology o.*

A policy statement only captures the semantics of the sentences that describe data collection, sharing or use. Other policy sentences that do not specify explicit data practices, such as "we will stop showing targeted advertising", are not modeled since it is unclear which data is collected or used. The data usage *du* can be a special value *None*, indicating that the usage purpose is not specified.

5.2 Opt-out Policy Classes

To analyze the (in)consistencies between opt-out policies and data practices, we categorize policy statements according to their stated data practices and purposes. Inspired by prior analysis of online trackers' policies [28, 66], we consider 5 types of opt-out policies: no tracking (*No-tracking*), no data collection via third-party cookies (*No-data-collection*), no data collection for targeted advertising purposes (*No-data-coll.-for-oba*), no displaying online behavioral advertising (*No-display-oba*) and *Other.* Since this work is only concerned with the transmission of cookies to tracker servers, we use the *No-data-collection* class to denote the termination of data collection *via third-party cookies*, rather than data collection with other means such as the IP protocol/addresses. Finally, the *Other* class includes samples that do not belong to any other classes, such as the opt-out of the sale of personal information or marketing communication, and opt-out instructions.

Our opt-out policy taxonomy covers two main types of data practices, *user-activity tracking* and *user-data collection*, while a data practice's purpose is either *for delivering OBA* or *unspecified.* However, the *No-tracking* class is not divided further based on the data-usage purpose because a statement about tracking is seldom coupled with data-usage purposes.

We formalize the opt-out policy classes in such a way that each policy class comprises policy statements that have semantically equivalent terms. For example, *No-tracking* class is a set of policy statements in the form *(r, collect, id_data), (id_data, not_for, tracking)* where *id_data* can be substituted by a synonym such as "unique identifier" and *r* can be a synonym of "first party". Of the opt-out policy classes, statements about stopping displaying OBA (*No-display-oba*) are not formalized for the flow-to-policy consistency analysis because they do not explicitly express any data collection. Table 1 lists the privacy-statement sets per policy class.

5.3 Automated Opt-out Policy Classification

The extraction of opt-out policies from a policy sentence is formulated as a binary classification problem. For each opt-out policy class, we create a classifier that determines whether a sentence expresses the opt-out policy or not. As the result, a sentence may contain one or multiple opt-out policies. For example, "to opt out of our tracking and data collection, please click the button below" contains two policies: *No-tracking* and *No-data-collection.*

The rest of this section details the steps of the automated classification pipeline: identify, extract and classify opt-out policy clauses.

5.3.1 Opt-out Predicate Identification. The pipeline first extracts opt-out predicates (verbs) describing the actions that a user needs to take to opt out. The predicates' most common form is a verb with lemma *opt.* Also, `OptOutCheck` looks for nouns with lemma *opt* and traverses up the dependency tree to identify the action performed on the nouns. For example, given "if you do not want to see OBA, please *click* our opt out here," *opt* is a noun and *click* is extracted.

5.3.2 Opt-out Policy Clause Extraction. To extract the clauses that express the data-collection policies for an opted-out user, the system identifies the clauses that have one of the following grammatical roles with respect to an opt-out predicate: object, main clause, and adverbial clause. In an exceptional case when a sentence does not have any opt-out predicate, but its context is clearly about opt-out policies (e.g., the sentence is the label of an opt-out button), we treat the whole statement as an opt-out policy clause. Table 2 lists examples of the opt-out policy clauses and their roles in a sentence.

The system primarily extracts the opt-out policy clauses from a sentence by analyzing the semantic arguments of the *opt-out* predicates. Specifically, we design `OptOutCheck` to analyze the following arguments of each opt-out predicate: object (*Arg1*), instrument (*Arg2*), adverbial (*Argm-Adv*), purpose (*Argm-Prp*) and purpose-not-cause (*Argm-Pnc*). A semantic argument answers questions like "who?", "did what?", "to whom?", and "for which purpose?" of an event expressed by the predicate [64, 68]. The definitions of these arguments are given in the OntoNotes 5 linguistic corpus [90].

As a complement to the semantic-role analysis, `OptOutCheck` analyzes the syntactic dependencies of *opt-out* verbs. In particular, it searches for the main clause of each *opt-out* predicate by analyzing the sentence's syntactic dependency tree [64]. For example, the verb *opt* in "if you opt out" does not have any semantic arguments, so `OptOutCheck` looks for its main clause "we will no longer use cookies to collect your data" and treats it as an opt-out policy clause.

Grammatical Role	Example	Policy Statement	Policy Class
Object	You can opt out of tracking and our unique cookie identifiers here.	(we, collect, data), (data, not_for, tracking)	*No-tracking*
Main clause	If you opt out, we will no longer use cookies to collect your data for targeted advertising.	(we, collect, data), (data, not_for, targeted ad)	*No-data-coll.-for-oba*
Adverbial clause	If you want us to stop collecting your data, please opt out here.	(we, not_collect, data), None	*No-data-collection*
No "opt" predicate	Please do NOT collect information about me using cookies and other tracking technologies.	(we_implicit, not_collect, data), None	*No-data-collection*

Table 2: Examples of opt-out policy clauses, their grammatical roles with respect to the *opt* predicate, the extracted policy statements and opt-out policy classes. The opt-out policy clauses in each sentence are underlined.

5.3.3 Opt-out Clause Analysis. OptOutCheck classifies a sentence into the opt-out policy classes by identifying data objects, data-collection sentiment (i.e., collect or not) and advertising data-usage purposes in an opt-out policy clause. To identify the *No-data-collection* policy for "opt out of" phrases, OptOutCheck identifies negative data-collection actions on data objects in the object argument *Arg1* of an *opt* predicate. It uses a named entity recognition (NER) model [64] to accurately extract data objects (e.g., *cookies* and *unique cookie identifiers*). In addition, we use patterns of syntactic dependencies to identify data-practice noun phrases. Data-collection noun phrases such as "use of cookie" and "collection of data" are identified by searching for data objects (e.g., *cookie* and *data*) with a *pobj* (object of a preposition) dependency with respect to data-usage actions (e.g., "use" and "collection"). For example, "opt out of unique cookie identifiers" and "opt out of our use of information about you" are classified as *No-tracking* and *No-data-collection*, respectively.

Since cookies are the means of data collection, a negative-sentiment action performed on cookies is an indication of the *No-data-collection* policy, such as "we will stop placing cookies on your browser." The common actions on cookies are *drop*, *place*, and *set*. The negative sentiment of a data-collection action is indicated by the existence of a negation-modifier dependency, an *Argm-Neg* semantic argument, or a negative-sentiment modifier such as "no longer" and "stop".

Since the sentences in close proximity to opt-out buttons have a context related to opt-out choices, the occurrence of certain keywords is a good indicator of policy classes. Specifically, to extract *No-tracking*, the classifier looks for nouns and verbs related to tracking, such as *tracking*, *identifier* and *disassociate*. Similarly, advertising-related keywords, such as *target*, *advertising* and *marketing*, indicate advertising data-usage purposes. The advertising purposes also distinguish *No-data-collection* from *No-data-coll.-for-oba*.

5.4 Development of Opt-out Policy Classifiers

We create a manually-annotated dataset as the ground truth to develop matching patterns for the opt-out policy classifiers as follows.

5.4.1 Tracker Selection. We crawled cookies of the top 5k websites in the US as of October 2020, ranked by the SimilarWeb analytics service [74]. This selection is to ensure that the privacy policies of the online trackers are subject to the same legal and regulatory requirements, such as the Notice and Choice framework [23]. Moreover, we excluded pornography sites, using a blocking list [86], since they use specialized trackers [104] and are not our focus.

We selected a dataset of 120 popular third-party cookie domains. From the 180 cookie domains that were present on at least 100 websites, we chose the top 100 third-party cookie domains and other 20 randomly selected domains from the remaining cookie domains to cover both the most popular and less popular cookie domains. The number of domains was limited by the resources needed to analyze and annotate the cookie domains. Appendix C.1 provides details of the cookie collection and domain selection.

5.4.2 Opt-out Button Identification. From the selected cookie domains, we traced back to the websites of the trackers that own the cookie domains and manually extracted the opt-out buttons on each website. From the home page, we searched for the privacy policies (e.g., for website visitors, corporate customers, and end-users) and then identified the opt-out settings contained in the policies. Since opt-out buttons were not ambiguous, this extraction was done by one advanced PhD student and took an average of 45 minutes for each domain, or 90 hours for 120 cookie domains. Appendix C.2 provides details of the extraction process.

Of the analyzed trackers, 80 provided opt-out choices. The most common form is single-click opt-out buttons. 76 (95.00%) of the settings have a single step, i.e., a single click, to opt out. The remaining need 2 steps: select an opt-out preference option and click *submit*.

5.4.3 Opt-out Policy Corpus. From the identified opt-out web pages, we selected the sentences next to the opt-out buttons and classified them into the opt-out classes. Since privacy policy sentences were vague and complex, the classification was done by two PhD students with no less than 3 years of experience in user-privacy research. It took an average of 3 minutes for each sentence on average, or 20 hours for both annotators. The inter-annotator agreement is 94%. We held a follow-up meeting to reconcile the differences.

The final opt-out policy corpus contains 246 sentences in 80 trackers. *No-display-oba* is the most common opt-out policy with 49 (19.92%) occurrences. *No-data-collection* constitutes 23 (9.35%) instances. The least common policy with 18 (7.32%) samples is *No-tracking*. 56% of the sentences did not state any opt-out policies and were classified as *Other*. These sentences (next to opt-out buttons) explained opt-out mechanisms (e.g., opt-out is browser-specific or browser's cookie-functionality must be enabled for opt-out). The number of sentences per opt-out policy class is listed in Table 3.

5.4.4 Automatic Classifiers. Using the dataset, we derived two classifiers for *No-tracking* and *No-data-collection* policies which are the only opt-out policies that can be verified by observing the behavior of the trackers on the client side. Other classes related to online-behavioral advertising purposes are hard to verify without knowing the processing purposes on the tracker servers.

The classifiers achieved an average F1 score of 86.04% with precision ≥ 88% on the policy corpus. The high inter-annotator agreement and the high F-1 scores demonstrate the consistency of the interpretation of the policy classes and the regularity of the sentence patterns. Due to the data sparsity, i.e., small numbers of samples per opt-out policy class, we use the dataset as a training set for developing the matching patterns while Section 9.2.2 will evaluate their performance as part of the consistency analysis pipeline. Table 7 (Appendix C.3) provides details of the classifiers' performance.

Policy Class	# Sentences
No-tracking	18 (7.32%)
No-data-collection	23 (9.35%)
No-data-coll.-for-oba	23 (9.35%)
No-display-oba	49 (19.92%)
Other	139 (56.50%)
Total	246 (100%)

Table 3: Opt-out policy dataset. A sentence may contain multiple opt-out policies.

Metric	Train	Test
Precision	0.98	0.97
Recall	0.74	0.74
F1	0.84	0.84
Support	649	279
# Samples	7,649	3,279

Table 4: Opt-out cookie classifier performance on the training and test sets.

5.5 Implementation

5.5.1 Opt-out Policy Statement Identification. OptOutCheck extracts opt-out policies from the policy statements that describe the data collection practices after a user clicks on the opt-out button. For example, as shown in Fig. 1, advertisers would cease their tracking after the user opts out. Identifying these sentences is challenging because of the flexible design and implementation of websites.

We observe that the opt-out policy statements are commonly placed nearby (e.g., in the surrounding paragraphs). This assumption is close to the expectation of FTC [23]. Therefore, given an opt-out page identified in Section 4, OptOutCheck converts the web page into plain text [89, 94] and extracts 10 sentences (5 before and 5 after) surrounding the position of the opt-out button. Furthermore, to reduce unrelated statements, except for labels of opt-out buttons, policy sentences without any "opt" predicate (e.g., *opt-out*, *opt out* and *opting out*) are excluded.

5.5.2 Natural Language Analysis. OptOutCheck uses the neural-network-based language pipelines of the Spacy NLP library [4, 72] to parse and create the dependency trees of privacy policy sentences. The semantic arguments are analyzed by using a semantic role labeling model (SRL) [6] based on *Roberta-base* contextualized word embeddings and trained on the CoNLL2012 (OntoNotes 5) large-scale natural language dataset [96]. Finally, we use PurPliance [18] to analyze privacy-statement parameters such as data-collection actions and data objects. To improve the data-type extraction, we augment its data-object NER model with terms related to cookies that are commonly used in the privacy policies of online trackers.

6 OPT-OUT COOKIE EXTRACTION

To check whether a tracker's data collection practices follow its opt-out policies or not, it is necessary to determine that a user's opt-out preference has been recorded by the tracker. Since we focus on the opt-out mechanism based on anonymous cookies, we define *opt-out cookies* as the cookies that online trackers use to record a user's opt-out choice [5, 27, 38]. These cookies are created upon clicking an opt-out button for the trackers to enforce their opt-out data collection policies on web pages where the cookies are present.

Automated extraction of opt-out cookies is necessary as privacy policies rarely include specifications of these kinds of cookies. The mapping from a tracker to cookie domains using a predefined list is also not guaranteed to be complete and up-to-date. Furthermore, a differential analysis of the cookies before and after an opt-out is not sufficient for extracting opt-out cookies because the opt-out button may redirect the user to the tracker's home page where other cookies — unrelated to opt-out cookies — are added.

6.1 Opt-out Cookie Classifier

OptOutCheck takes a hybrid approach to extract opt-out cookies where a cookie is matched with a predefined opt-out cookie registry and then an automatic classifier if not found. The exact-match approach leverages the opt-out cookie registries provided by automatic opt-out tools: Evidon Global Opt-out [29], DAA Protect My Choice [7], and Google Keep My Opt-Outs [46]. Any cookie that has its name, domain and value matching the registries is determined as an opt-out cookie. The extraction excludes *session cookies* because the tracker should remember the opt-out choices of users over multiple browsing sessions. In what follows, we describe a classifier that uses the pattern of a cookie's name and value to determine whether it is an opt-out cookie or not.

6.1.1 Opt-out Cookie Dataset. To develop and evaluate the opt-out cookie matching patterns, we derive a ground-truth dataset that contains the cookie names and values from the exact-match registries. We excluded cookies with a non-anonymous identifier value, which is empirically identified as a combination of 10–20 alpha-numeric characters, while keeping cookies with anonymous values that comprise only zeros and dashes. This process resulted in 928 opt-out cookies from 795 trackers.

We then mixed the opt-out cookies with 10k cookies randomly sampled from the crawling of the top 5k websites as described in Section 5.4.1. These additional cookies are considered negative samples (i.e., non-opt-out cookies) because the crawling process did not perform any opt-out, i.e., we assume the browser does not have any opt-out cookies unless the user explicitly opts out.

Stratified partitioning was then performed to split the dataset into training and test sets with a 70–30% ratio. The patterns are developed on the training set and evaluated on the test set. The final dataset contains 10,928 cookies with 7,649 and 3,279 samples in training and test sets, respectively. The number of samples and supports in the dataset are shown in Table 4.

6.1.2 Opt-out Cookie Patterns. The matching rules comprise two types of patterns based on cookie names and cookie values. First, the patterns in cookie names include the spelling and abbreviation variants of "opt out", such as "opt-out" and "OptedOut". The abbreviation pattern "oo" does not simply match when it is a substring; it matches only if "oo" is either the whole string or surrounded by delimiters like "_". We exclude the cookies whose string values can be converted to *False* in common programming languages, such as *0* or *false*. For example, cookie *optout=false* does not indicate an opt-out. Second, a cookie is considered for an opt-out purpose if its name indicates a unique user ID, such as "uid" and "uuid", *and* its value is not unique such as a single-digit number like "-1" or "nan". These special values of a tracking cookie can be used to indicate the opt-out preference. It is worth noting that opt-out cookies must have both appropriate *key* and *value*, e.g., a cookie named "uuid" is not an opt-out cookie until its value becomes "-1".

6.1.3 Performance Evaluation. As shown in Table 4, the classifier achieves a high F1 score of 84% (97% precision and 74% recall) on the test set. As the dataset is highly unbalanced, these metrics are computed only for positive samples. We aim to minimize the false detections (i.e., maximize precision), so we consider the performance is good enough when the precision on the training set was greater than 95%. We conjecture that this high accuracy comes from the regularity of the naming of opt-out cookies created by programmers. It is worth noting that OptOutCheck does not recognize cookies with obfuscated names and values but this limitation does not increase the false-positive rate of the system.

7 DATA FLOW ANALYSIS

We now describe how OptOutCheck extracts the actual data-collection behavior of a tracker from its network traffic to detect the inconsistencies, if any, between its actual behavior and opt-out policies.

7.1 Data Flow Definition

We consider the data objects and purposes in the data-collection behavior of a tracker, which is formalized as follows.

Definition 7.1 (Data Flow). *A data flow is a 3-tuple (r, d, p) where a recipient r collects a data object d for the receiver's purpose p.*

The receivers of network traffic are determined by the destination hosts in the intercepted URLs. For example, the data sent to hosts owned by tracker T has the receiver $r = T$. A data object d is the data type transferred via the network, such as "unique identifier" or "user location". A data-usage purpose p is the purpose of collecting and using the data object such as "for delivering OBA" or "for product research and analytics."

7.2 Extraction of Key-Values

In order to extract key–value data pairs from cookies and URL parameters in the HTTP traffic, OptOutCheck addresses two challenges: 1) ensure captured traffic falls under the scopes of the corresponding opt-out policies and 2) avoid cookies that are only stored in the browser but not transferred to the servers.

7.2.1 Opt-out Policy Scopes. To analyze the data collection on opt-out choices, OptOutCheck considers only cookies and URL parameters sent to the URLs that fall under the scope of opt-out policies. In particular, these URLs are the ones that match the domains of the tracker's opt-out cookies (determined in Section 6). Although the scope of opt-out choices may span beyond the opt-out cookies' domains, because a tracker must own the domain of an opt-out cookie, we assume a data flow to follow the opt-out policy if its domain matches the top-level domain of an opt-out cookie, called an *opt-out domain*. For example, if the opt-out cookie is *opt_out=1* under domain *ads.tracker.com*, the opt-out domain is *.tracker.com*. The domain matching follows the domain-match specification [13]. Moreover, the longest matching URL paths take precedence if there are multiple matched domains and paths found [30].

7.2.2 Cookie Transfer Interception. OptOutCheck intercepts the cookies and URL parameters transferred from a web browser to the trackers' servers in the HTTP requests made by the browser during each web page visit. By capturing the cookies transferred via network traffic, the data in the cookies is guaranteed to be collected by the trackers, rather than being only stored and unused in the browser. To determine the expiration time of the cookies intercepted in the HTTP requests which contain only the keys and values of the transferred cookies, they are resolved to the cookies stored in the browser by matching their names, values, domains, paths and request URLs. We use the HTTP request interception feature of the web browser automation tool where the interception is performed before the traffic is encrypted in the HTTPS protocol.

7.3 Extraction of Data Flows

From the extracted key–value pairs, OptOutCheck infers the data objects d and data-usage purposes p of data-flow tuples formalized in Definition 7.1. For example, a data flow associated with the collection of a unique-ID cookie *uid* used by a tracker T is $(T, uid, tracking)$. Since the automatic opt-out policy extractors extract only *No-tracking* and *No-data-collection* opt-out policies (Section 5.4.4), we focus on detecting the data types that reflect the tracking and data collection of a tracker as follows.

7.3.1 Detection of Tracking Identifiers. OptOutCheck detects the cookies that contain unique identifiers for tracking purposes. A data flow for such a tracking cookie is *(<tracker>, unique ID, tracking)* where *<tracker>* is the tracking cookie's owner. Unique IDs (known as unique user identifiers or tracking IDs) are widely used for tracking users [21, 87, 93].

Since automatic detection of identifier cookies has been developed before [36, 45, 77], we assume cookies and URL parameters containing unique IDs are used for tracking purposes. While it is not possible to determine the ultimate usage purposes of these IDs without the information on the server side, unlike automatic data collection such as logging of IP addresses on HTTP servers, setting cookies and URL parameters requires significant effort, and hence the collection of such data is unlikely to be accidental. For example, the collection of a cookie named *uid* containing a 16-digit identifier that does not change throughout a user's browsing activity is likely to track users by assigning each user a unique user ID.

OptOutCheck determines a cookie to have a unique ID using a set of criteria that are empirically determined and evaluated by Englehardt et al. [36]. The heuristics leverage two main properties of a unique ID cookie — *unique across browser instances* and *persistent over time*. There are 5 criteria as follows. First, cookies are *long-lived*, i.e., their expiration time is longer than three months. This time threshold is the same as that in the work of Englehardt et al. [36]. Second, their values are *constant* throughout web browsing (i.e., visits to different websites by the same browser instance) to avoid varying non-ID values like timestamps and browsing history. Third, the cookie values are of *constant length* across different measurements. Fourth, cookies have *user-specific* values which are unique among different browser instances. Finally, cookie values have *high entropies*, i.e., their values change significantly across measurements. A cookie is filtered out if the RatcliffObershelp-similarity [88] score of its values in different measurements is higher than 0.55. Note that OptOutCheck reuses the threshold values from [36] and developing better thresholds is outside of this paper's scope.

OptOutCheck parses and decodes URL parameters into key–value pairs in order to determine the data types collected by the trackers. As the values can be encoded in various data formats [21], OptOutCheck attempts to decode the URL parameters and cookie values in JSON and base64 formats. The same heuristics of detecting unique IDs for cookies apply to URL parameters except for the long-lived criterion as URLs do not have any expiration time.

7.3.2 Detection of General Data Collection. In addition to the unique IDs, `OptOutCheck` detects the collection of other user data types such as location and web browsing history. Inspired by the *bait* technique [1], the system looks for the known values of the crawling servers' IP addresses, location (e.g., city and state names), browser/OS versions, and URLs of the visited web pages in the values of the extracted key–value pairs. Their existence is the indication of data collection by a tracker. For example, a tracker is collecting user location if its cookie contains a key–value pair *region=<city_name>* containing the name of the city where the crawling server is located.

8 OPT-OUT FLOW-TO-POLICY CONSISTENCY

This section presents a formal model to analyze the consistency between the policy statements and data flows.

8.1 Subsumptive Relationship

The formal representations of opt-out policy statements and data flows (Definitions 5.1 and 7.1) are based on the concepts of receiving entities (i.e., receivers), data objects and purposes that have subsumptive relationships with each other. For example, a relation "personal data includes email addresses" translates to that *email address* is subsumed by *personal data.* `OptOutCheck` leverages the subsumptive relationships in PolicyLint ontologies [9] that are derived from subsumptive phrases of a large number of privacy policies. The policy terms' relationship is formalized as follows.

Definition 8.1 (Subsumptive Relationship). *Concept x is subsumed by another concept y, denoted as $x \sqsubset_o y$, if and only if $x \not\equiv_o y$ and there is a path from y to x in an ontology o represented as a directed graph in which each node is a term and each edge points from a general term y to a specific term x included in y, i.e., x "is a" instance of y. Similarly, $x \sqsubseteq_o y \Leftrightarrow x \sqsubset_o y \vee x \equiv_o y$.*

8.2 Consistency Model

Informally, a data flow is consistent with a privacy policy T which consists of a set of policy statements t_s, if there is a policy statement that discloses the data object and purpose of the data flow and there is no policy statement that discloses otherwise (e.g., uncollection of the data). The consistency condition is formalized as follows.

Definition 8.2 (Flow-relevant Policy Statements). *A privacy statement $t_f = ((r_t, c_t, d_t), (e_t, k_t, q_t))$ is relevant to a flow $f = (r, d, p)$ (denoted as $t_f \simeq f$) iff $r \sqsubseteq_\rho r_t \wedge d \sqsubseteq_\delta d_t \wedge p \sqsubseteq_\kappa p_t$. Let T_f be the set of flow-f-relevant policy statements in the set of policy statements T of a privacy policy, then $T_f = \{t_f \mid t_f \in T \wedge t_f \simeq f\}$.*

Definition 8.3 (Flow-to-Policy Consistency). *A flow f is said to be consistent with a privacy policy T iff $\exists t_f \in T_f$ such that $c_t = collect \wedge k_t = for$ and $\nexists t'_f \in T_f$ such that $c'_t = not\_collect \vee k'_t = not\_for$.*

A data flow is inconsistent with a privacy policy if the Flow-to-Policy Consistency condition is not satisfied. For example, an opt-out policy *((ad_platform, collect, data), (data, not_for, tracking))* is inconsistent with a data flow *(ad_platform, user_ID, tracking)* when the ad platform still retains a user ID cookie *uid=<unique_ID>* to track users after an opt-out even though the policy states that they will cease their tracking practice. For the sake of brevity, the definitions are for policy statements with a specified usage purpose. If the data usage purpose du of a policy statement is unspecified, i.e., $du = None$, the conditions on the data usage purpose are ignored.

8.3 Inconsistency-Detection Rules

`OptOutCheck` detects two types of consistency corresponding to the two opt-out policy classes. If the opt-out policy is *No-tracking*, the collection of unique IDs for tracking purposes after the user opted out is inconsistent. If the policy is *No-data-collection*, the collection of any data (such as unique IDs, user location, web page URLs and IP addresses) is inconsistent. The following theorem formalizes an inconsistency when a tracker still collects unique IDs for tracking purposes after users' opt-out. Appendix D shows the proof.

Theorem 8.4 (Unique-ID Tracking Inconsistency). *The collection of unique IDs for tracking purposes after users' opt-out is inconsistent with a No-tracking or No-data-collection opt-out policy.*

9 LARGE-SCALE STUDY

9.1 Tracker Selection

We selected widely-used tracker lists that provide the websites of trackers' owner companies and privacy policies to derive a tracker dataset. In particular, we used 4 tracker databases: WhoTracksMe [44, 65], Disconnect Tracking Protection [32], Evidon Global Opt-out [29] and DuckDuckGo Tracker Radar [33]. These databases had 3,194, 1,393, 796 and 229 trackers, respectively. Table 8 (Appendix E) provides the details of the trackers in these databases. We did not use the tracker domains in ad-blocking lists because many of them were resolved to only file servers without obvious connections to the trackers' privacy policies [34].

By uniquely identifying each tracker by its pay-level domain, merging the three selected lists yielded 4,021 unique trackers. The number of trackers that the crawler successfully loaded a home page is 3,319. Finally, we removed trackers with home pages redirected to the same website domains, leaving 2,981 trackers. This step is to avoid those ad platforms that provide multiple different ad services. For example, 29 home pages of Google ad services had the same *google.com* domain. We did not exclude non-English home pages at this stage to avoid removing multilingual trackers which might have a non-English home page but an English privacy policy. Fig. 4 depicts the numbers of trackers filtered by each selection step.

9.2 Extraction of Opt-out Buttons and Policies

9.2.1 Opt-out Button Extraction. From the selected 2,981 trackers, the Google Programmable Search Engine yielded 14,059 links for 71.72% (2,138/2,981) trackers. Only 62 (2.1%) of the tracker websites disallowed the Google search engine by using *robots.txt.* Refining the search results to only the top-3 links and removing links to PDF files (e.g., PDF privacy policies) yielded 5,323 links for opt-out page candidates of 71.05% (2,118/2,981) trackers.

Extracting opt-out buttons from the opt-out page candidates led to opting out 195 trackers, i.e., detected an opt-out button and found opt-out cookies after clicking the button. After excluding 30 trackers with non-English opt-out pages, `OptOutCheck` identified 265 opt-out cookies from 165 trackers. Using only the pattern-based classifier, it could still identify 254 opt-out cookies from 160 trackers, demonstrating the effectiveness of opt-out-cookie patterns. Fig. 4 depicts the trackers throughout the opt-out detection process.

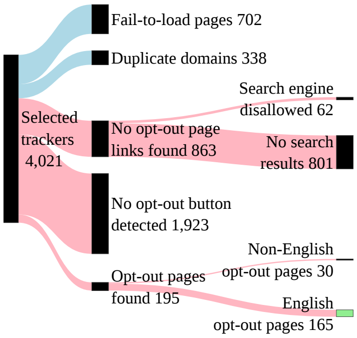

Selected trackers 4,021
Fail-to-load pages 702
Duplicate domains 338
No opt-out page links found 863
Search engine disallowed 62
No search results 801
No opt-out button detected 1,923
Opt-out pages found 195
Non-English opt-out pages 30
English opt-out pages 165

Figure 4: Opt-out page detection. Each label contains # of trackers. Blue lines: tracker selection. Red lines: opt-out detection. Green box: results for further analysis.

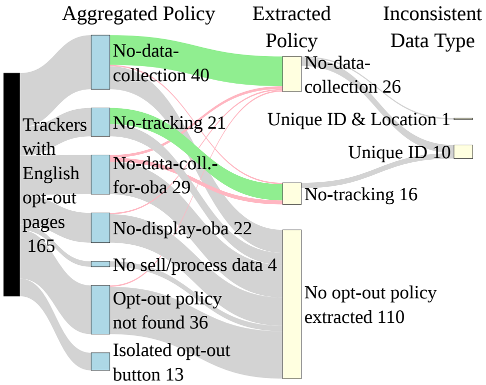


Figure 5: Opt-out policies and inconsistent data collection. A tracker's policies are aggregated to a single policy. Green/red lines: true/false positives. Blue/yellow boxes: manual/auto analysis.

Policy Class	# Sentences	# Trackers
No-tracking	30 (2.29%)	24 (15.79%)
No-data-collection	38 (2.91%)	34 (22.37%)
No-data-coll.-for-oba	41 (3.13%)	37 (24.34%)
No-display-oba	106 (8.10%)	43 (28.29%)
Other	1,171 (89.53%)	52 (34.21%)
Total	1,308	152

Table 5: Ground-truth opt-out policy dataset. A sentence or tracker may include more than 1 opt-out policy class.

Policy Class	Precision	Recall	Extracted # Sents. (# Trks.)
No-tracking	84.6%	73.3%	26 (20)
No-data-collection	85.2%	60.5%	27 (25)
Total			52 (40)

Table 6: Policy classifier performance on the ground-truth dataset.

9.2.2 Opt-out Policy Extraction. Of the 165 trackers with English opt-out pages, the system found 1,369 sentences related to opt-out policies in the privacy policies of 152 trackers. `OptOutCheck` extracted 55 *No-data-collection* and *No-tracking* policies from 54 sentences of 42 trackers. A tracker may contain multiple opt-out policies. *No-data-collection* is the most common policy class with 26 trackers. The rest only provided OBA opt-out. Fig. 5 depicts the policies extracted from the selected trackers.

Of the 49 trackers with no opt-out policies extracted (bottom of Fig. 5), 36 did not include any specific opt-out-policy statements in the 10 sentences next to an opt-out button (e.g., when the button was placed on a sidebar). 13 of the trackers embedded the opt-out button in an isolated *iframe* that contained only the opt-out button.

9.3 Detection Performance Evaluation

9.3.1 Opt-out Button Extraction. To evaluate the accuracy of the opt-out button extractor, we randomly selected 50 trackers in the tracker dataset (Section 9.1) and manually identified the opt-out choices provided by these trackers. The sample trackers were unseen during the development of the extractor's button matching patterns (Section 4.2). Of these, we found 10 trackers providing opt-out buttons (the other 4 trackers were excluded because their opt-out buttons led to nonexistent web pages or the policies were not written in English). The opt-out button extractor extracted 5 buttons with a precision of 100% and a recall rate of 50%.

The number of the detected opt-out buttons was not high because the trackers either provided no opt-out choice or used a complex (or difficult-to-find) opt-out process. Of the 40 trackers without opt-out buttons: 9 had non-English privacy policies; 8 were placeholder (e.g., domain-for-sale) sites; 17 had no concrete opt-out instructions; and 6 instructed users to either contact them via email/form-submission, or use opt-out buttons in emails. The latter 23 trackers might have less incentive to implement a cookie-based opt-out because they either 1) provided services other than website-based advertising (e.g., email marketing (*selligent.com*), marketing-automation (*activedemand.com*), or CDN (*akamai.com*)); or 2) were trackers with small market shares.

9.3.2 Opt-out Policy Classifier. To evaluate the performance of the opt-out policy classifier, we create an unbiased ground-truth opt-out policy dataset by annotating 1,308 sentences extracted from the tracker privacy policies (Section 9.2.2) that were unseen by the classifier during the training phase (Section 5.4.3). Specifically, two authors independently annotated these sentences into the 5 policy classes using the same procedure as in Section 5.4.3. Each sentence took 1.5 minutes to classify on average, or 65.4 hours for the two annotators to complete. The inter-annotator agreement was 96%.

As shown in Table 5, the relative distribution of the policy classes is the same as the opt-out policy corpus (Section 5.4.3). The percentage of the Other class is higher than that in the opt-out policy corpus as the automated tool included more sentences next to an opt-out button than those manually selected in the policy corpus.

Our results show that the classifiers achieved high precision rates of 84.62% (22/26) and 85.19% (23/27) for the *No-tracking* and *No-data-collection* classes, respectively. Since we aim to minimize the false positives, the recall is not as high as the precision. The recall rates of *No-tracking* and *No-data-collection* are 73.33% (22/30) and 60.53% (23/38), respectively. Table 6 shows the policy classification results. Note that achieving a high recall while maintaining a high precision is still an unsolved problem for existing privacy-policy analysis techniques [9, 15, 18].

A few false positives still occurred due to `OptOutCheck`'s failure to identify the advertising scope of an opt-out, such as when a statement referred to an ad-delivery purpose in a previous sentence. For example, *admedo.com* stated that the opt-out would prevent them from collecting "data from your browser for *these purposes*" while "these purposes" referred to the delivery of their OBA, so the opt-out class is *No-data-coll.-for-oba* instead of *No-data-collection*.

The false negatives were due to the limitation of the sentence-based analysis and the small number of the extraction rules which failed to handle complex grammar of legal statements. Of the 15 missed *No-data-collection* statements, 3 sentences did not include any "opt" predicate while referring to an opt-out method in a previous sentence. For example, "by using this tool, a third-party cookie will be added to your browser so that you will no longer see Yieldify campaigns or have data collected by Yieldify," where OptOutCheck was unable to resolve "this tool" to the opt-out tool. Furthermore, the rule-based extractor utilized precise but inflexible matching rules while the small training set limited the number of rules (Section 5.4.3). For example, 3 *No-data-collection* sentences included a statement pattern "if you *do not want us to* collect your data, you can click on the opt-out button" while this pattern was unseen by the classifier during the training phase. On the other hand, the higher recall of the *No-tracking* opt-out policies indicates a lower variation of the grammar of the *No-tracking* statements.

Fig. 5 visualizes the high precision of the policy classifier. It shows the opt-out policies extracted by OptOutCheck and manually identified by the authors, based on the annotations of all 152 trackers in *both* Sections 5.4.3 and 9.3.2. The thickness of the green and red lines illustrate the true and false positives of the opt-out policy classification, respectively. To avoid being cluttered with numerous combinations of different opt-out policies that a tracker may have, we aggregate a tracker's policy classes into an *effective* opt-out policy used to detect the inconsistencies. The precedence order we used is *No-data-collection*, *No-tracking*, *No-data-coll.-for-oba* and *No-display-oba*. For example, if a tracker states both *No-data-collection* and *No-tracking*, it is classified as *No-data-collection* which is a superset of *No-tracking*. Furthermore, we order OBA-related policies later and separate opt-out policies related to data selling/processing since they cannot be verified given only the client-side information.

9.3.3 Analysis of "Other" Statements. We further classified the statements labeled with "Other" to understand the contextual statements of opt-out choices. Two authors read the "Other" statements and discussed to create a taxonomy of 6 high-level categories. The most common class includes the explanation of opt-out mechanisms, (such as "Our opt-out tool is cookie-based") and step-by-step instructions (such as "Click the link below to set or remove the opt-out cookie"). Table 9 (Appendix F) provides the number of statements of the most common categories.

9.4 Data Flows and Opt-out Inconsistencies

9.4.1 Measurement Procedure. We analyze the differences in cookies on publisher websites between before and after opting out of a tracker T to detect the changes in the data-collection behavior of T. This process avoids false positives due to the cookies set by the tracker's own website when OptOutCheck visited it for opting out. These cookies may entail first-party data collection of T that is unrelated to T's third-party tracking services. Specifically, OptOutCheck first visits a set of publisher websites using a clean instance of a web browser and records the set T_c of cookies under T's opt-out domains. It then visits the tracker's website and activates the opt-out choices provided by T. OptOutCheck confirms that the opt-out has been set successfully by checking the presence of the tracker's opt-out cookies. Finally, OptOutCheck visits publisher websites again and records the values of the cookies in T_c.

Due to the randomness of placement of online advertisements, OptOutCheck sequentially visits a set of candidate web pages S until it finds 10 web pages that send requests containing the cookies of T, or S is exhausted. We use the *WhoTracksMe* and DuckDuckGo Tracker Radar cookie databases that contain the lists of trackers detected on top websites to generate S for each tracker.

9.4.2 Extracted Data Flows. Of the 165 trackers with opt-out buttons, OptOutCheck found 129,286 candidate websites for 146 trackers where their cookies may have been placed. Each tracker has an average of 582 (*SD* 1,026) candidate websites.

Following the measurement procedure in Section 9.4.1, OptOutCheck scanned 476 websites and extracted 52 data flows from 4,341 cookies for 33 trackers. Unique identifiers are the most common data type and are found on 98% of the flows. The other data type is the information about the user's IP address and city name included in the cookie *geode* of *udmserve.com*.

9.4.3 Detected Inconsistencies. OptOutCheck detected 11 trackers that had conducted tracking and data collection inconsistently with their opt-out policies after activating the opt-out choices. Fig. 5 depicts the distribution of the detected inconsistent data types.

Although the number of the detected inconsistent trackers is low, they tracked a significant amount of web traffic while the inconsistencies are direct violations of the trackers' privacy policies. On average, each tracker was present at 0.64% (*SD* 1.27%) across all page loads and on 3.65% (*SD* 6.57%) of the top 10k websites where they were included as a third party in March 2022 [44]. Given that there were 4.95 billion Internet users [100], these inconsistencies might affect a significant number of users.

Opt-out Domains. Most of the policies lacked the specification of their opt-out domains, highlighting the need of analyzing opt-out scopes (Section 7.2.1). Of the 11 inconsistent trackers, none specified the domains to which opt-out policies apply, 5 only stated the placement of opt-out cookies and 1 specified the opt-out cookie's domain (*adtriba.com*, Fig. 1). Moreover, 5 trackers had opt-out-cookie domains mismatching their policy URLs (Table 13).

9.4.4 Validation of Detected Inconsistencies. Two authors independently verified the results by manually following the measurement procedure (Section 9.4.1) and checking the existence of tracking cookies using Chrome DevTools. We determined the purposes of cookies from cookie names, values and cookie description (if there is any). All of the detected inconsistencies were confirmed to be correct (i.e., the detection precision is 100%). Appendix G provides details of the detected inconsistent flows, opt-out policies, opt-out cookies and domains.

9.4.5 Recall Evaluation. To estimate the recall of OptOutCheck's inconsistency detection, we selected 61 trackers with either *No-data-collection* or *No-tracking* policies from the 152 trackers with opt-out policy annotations in Sections 5.4.3 and 9.3.2. We then attempted to manually opt out of the trackers by following Section 9.4.1.

We found 16/61 (26.23%) trackers to have data flows inconsistent with their opt-out policies, yielding a recall rate of 68.75% (11/16). The opt-out policy classification was one of the main bottlenecks of the detection pipeline. While OptOutCheck correctly identified the opt-out cookies and extracted data flows from the trackers, the detection missed some inconsistencies due to the false negatives of the opt-out policy classifier. Furthermore, finding appropriate publisher websites to extract trackers' data flows was also a factor that limited the number of detected trackers. 8/61 (13.11%) trackers had either none or only 1 candidate web page (potentially due to their small market shares), making OptOutCheck unable to find any publisher websites that placed the trackers' cookies.

9.4.6 Case Studies. Criteo, which was present on 21% of the top 10k websites [44], contains multiple statements describing how its opt-out choice works such as "disable Criteo services will result in the deletion of the cookies dropped by Criteo in your browser you are currently using that allows us to recognize your browser or device" and "the termination of the collection of your personal data." Therefore, the opt-out policies are *No-tracking* and *No-data-collection.* However, after clicking "disable Criteo services" and the opt-out cookie *optout=1* was set, cookie *uid* was still retained with a unique ID. Both of these cookies were under *.criteo.com* domain.

Underdog Media instructed users to "opt out of our Underdog Media hosted technology by clicking here." After clicking the opt-out button, the website confirmed the status of "opt-out for Underdog Media hosted 3rd Party Cookies." Therefore, this opt-out policy was classified as *No-data-collection.* The button set an opt-out cookie *optout=Thank_You* but the tracker still retained multiple cookies to collect data from users. One of the cookies was *geode* which contained the IP address and city name of the browser.

Similarly, *adtriba.com* stated that "to be excluded from Adtriba third party tracking, you can click the following button." This policy was classified as *No-tracking*. However, even with an opt-out cookie *atboptout=1*, users were still tracked. The tracker retained an *atbgdid* cookie that contained a device ID [42]. This cookie was under *.adtriba.com* domain and existed on publisher websites even before our visit to *adtriba.com* to opt out, so it was likely used for third-party tracking. However, given the *No-tracking* opt-out policy, we expect all tracking cookies to be removed after an opt-out.

9.4.7 Root Cause Analysis. The inconsistencies could be due to an incomplete/buggy implementation of opt-out choices since trackers might not always develop and test the feature completely. In all the detected inconsistencies, the opt-out cookies were successfully set after clicking the opt-out buttons, demonstrating that the trackers made an effort to record opt-out preferences. However, the tracking cookies were still retained, so we hypothesize that the trackers are not successful at making the opt-out choice fully functional.

Since trackers have incentives to keep users from opting out of their tracking, they might attempt to make the opt-out process unnecessarily complex for the end-users. 3 trackers in the detected inconsistencies did not automatically delete their tracking cookies. For example, *criteo.com* retained the *uid* cookie after opt-out although the cookie did not reappear after its deletion. However, since many trackers automatically deleted their tracking cookies upon opt-out, there should be no difficulty in the automatic deletion of trackers' own cookies. Therefore, it is unreasonable to require average end-users to open Chrome DevTools to manually search and delete tracking cookies while retaining necessary opt-out cookies.

Regardless of whether the inconsistencies were accidental bugs or deliberately created by the trackers to mislead the users, since the opted-out users revoked their consent to tracking and/or data collection, the tracker companies conducted inconsistent data practices without the opted-out users' consent. Therefore, the companies may face heavy fines from regulators due to the deceptive privacy practices and unlawful data collection. It is a tracker's responsibility to ensure the consistency between its stated privacy policy and the actual data practices of its services. Given the detection of such inconsistencies by `OptOutCheck`, the trackers, developers and regulators can investigate and resolve their root causes.

9.5 Notification to Vendors

Of the 11 detected inconsistent trackers, we informed 10 trackers of the detected inconsistencies in their opt-out choices. We excluded *deepintent.com* because it made a drastic update on the website and removed the opt-out choice when we contacted them. Each of these notification emails included our interpretation of opt-out policies, our detected opt-out and tracking cookies, and the steps we took to reproduce the inconsistencies for each tracker. All of the notification emails appeared to have been delivered successfully. Fig. 6 (Appendix H) provides a template of the emails.

One of the trackers responded to our notification and subsequently made changes to its privacy policies to correct the detected inconsistencies. In particular, Taboola's Privacy Team confirmed our finding of their opt-out inconsistency. They then updated their opt-out method to immediately delete the tracking cookie *t_gid* after an opt-out, which also set an opt-out cookie DNT=1, to stop tracking users. They changed the opt-out button to point to a dedicated opt-out portal [101]. Specifically, they said "To avoid any confusion, and in an excess of caution, we have since updated the opt-out in our privacy policy so that it goes directly through Taboola's Data Subject Access Request Portal [101] instead and the user's t_gid cookie is deleted straight away." Two of the authors independently verified that their changes corrected the opt-out inconsistency. Taboola was categorized as "very prevalent" and ranked at 37/920 top most prevalent trackers on the Web while its cookies were present in 2.4% of all page loads on the top 10k websites [44].

10 LIMITATIONS AND FUTURE WORK

A major challenge in the analysis pipeline is the extraction of the sequence of user actions to activate an opt-out choice and the relevant policy statements from complex website content. Specifically, existing textual information extraction techniques are not applicable to extracting multi-step interactions and complex legal statements from general free-form website layouts. Although we have developed heuristics to extract policy statements from the sentences next to an opt-out button, a holistic analysis of the whole privacy-policy web pages will likely improve the recall rate. For example, *adform.com* placed the opt-out buttons on the sidebar far away from the opt-out policy statements, preventing/hindering `OptOutCheck`'s extraction of the opt-out policies. However, document-level analysis needs advances in natural language understanding and information extraction that have been studied extensively for decades [75].

`OptOutCheck` analyzes policies on a sentence basis and hence misses several cases due to the references to a previous sentence, such as in "we will stop this process when you opt out" where "this process" refers to the data collection for targeted advertising in the previous sentence. A holistic analysis of multiple sentences or the whole document will improve the recall rate. We did not check contradictions in opt-out privacy policies either, because the opt-out choice descriptions are usually short, and hence unlikely contain contradictions. Furthermore, the opt-out policy corpus in Section 5.4 is still small. We plan to use ML-based opt-out policy classifiers trained on a larger dataset to cover flexible grammar in privacy policies. This is part of our future inquiry.

We have not addressed other storage mechanisms (e.g., HTML5 LocalStorage), and advanced web tracking mechanisms (e.g., canvas fingerprinting [2]) due to the vagueness of their privacy policies. While the opt-out policies provided the definitions of cookie-based data collection and/or tracking, concrete descriptions of other technologies were often omitted. So, we leave the analysis of other tracking technologies as future work.

It is challenging to analyze data types and usage purposes of cookies without knowing their server-side processing. Unlike well-defined programming API (e.g., Android API), most cookies have no such specification of their purposes and value ranges. Furthermore, for security and performance reasons, the values of cookies are usually not human-readable but encrypted or encoded. Despite these challenges, researchers attempted to extract the purposes of transferred data from client-side information only [60, 97]. So, we leave the analysis of complete purposes of cookies as future work.

Major cookie-blocking desktop web browsers (e.g., Edge and Firefox) do not block *all* third-party cookies by default [79, 81]. Cookies used for certain purposes, such as analytics, are not blocked by Edge using the default browser settings. For example, we found empirically that Firefox and Edge still allowed *adnxs.com*'s tracking cookies on *cnn.com*. So OptOutCheck is still valid for these browsers.

Although our corpus focuses on websites in the US and privacy policies written in English, OptOutCheck is applicable to inconsistent trackers in other languages and countries. The implementation of opt-out choices and related policy statements may vary with the requirements of local privacy laws. However, analyzing the differences in the countries' regulations is outside this paper's scope.

While browsers may stop supporting third-party cookies in the future, OptOutCheck still applies to new tracking technologies. Analyzing them does not affect OptOutCheck's inconsistency-detection rules once data flows and policies are represented as formal tuples, and requires only key–value extraction from data storage (e.g., Local-Storage and IndexedDB [26]). Since policy structures do not change much over time [8], the policy analysis only needs the addition of new tracking-technology keywords. Furthermore, Google Chrome has continually deferred the blocking of third-party cookies with a plan of supporting them until 2024 [69].

Automatic detection of the discrepancies between the stated privacy policies and the actual data-collection behavior of trackers benefits all stakeholders of the Web ecosystem. First, regulators can readily scan trackers for critical violations to protect users. Second, the end-to-end automated framework can be easily integrated into the workflow of companies to assess the potential privacy risks in their system and gain more trust from users. Finally, users can avoid privacy risks due to misleading policy statements. We have already set up a website at [17] to increase user awareness of the online trackers' inconsistent opt-out choices discovered so far.

11 RELATED WORK

While there is research on cookie consent settings and opt-out choices in the privacy policies of publisher websites [12, 50, 51, 76], OptOutCheck's scope of online trackers is very different. Likewise, prior work on flow-to-policy consistencies of Android apps [9, 10, 18] does not directly apply to online services. We summarize prior research on online trackers and their opt-out policies.

Opt-out Choices of Online Trackers. Balebako *et al.* [11] measured the effectiveness of privacy tools including the opt-out cookie mechanism and found that the opt-out cookies were effective in limiting OBA. Sakamoto *et al.* [92] studied the opt-out cookie mechanism provided by ad agencies for opting out of OBA to find that the advertisers continued to track users when the users started to browse again. However, these are limited to evaluating the effectiveness of opt-out tools without systematically considering opt-out policies such as *No-tracking* and *No-data-collection* for opted-out users.

Komanduri *et al.* [66] examined the privacy policies of members of DAA and NAI to evaluate their compliance with the self-regulatory principles on the top 100 websites and reported non-compliance instances. They found that 93% of the 74 surveyed policies provided their own definitions of the opt-out and 57% provided opt-out definitions stronger than the minimum requirements of DAA and NAI. However, they assumed that the advertisers would honor the opt-out preferences. Cranor *et al.* [28] manually analyzed 75 privacy policies of advertisers who were members of DAA, and found the policies kept silent on many consumer-relevant practices. Our tools analyze the policies automatically and go beyond the members of DAA and NAI. Although these studies laid a foundation for analysis from a legal perspective, they did not develop any automated method to extract information from privacy policies.

Measurement of Online Trackers. Numerous researchers have studied the network of online trackers. Englehardt *et al.* [35] conducted large-scale measurements of online trackers on the top 1M websites. Lerner *et al.* [71] conducted longitudinal measurements of third-party web tracking for 10 years and found increasing prevalence and complexity of third-party tracking on the Web. Iordanou *et al.* [59] analyzed the data flows across the borders of EU nations and found that the majority of tracking flows cross countries in Europe but are well confined within the GDPR jurisdiction. Yang *et al.* [106] compared web tracker ecosystems on desktop and mobile environments. However, the prior work has not analyzed the trackers' privacy policies and verified whether the tracking practices followed the opt-out policies or not.

12 CONCLUSION

We have presented OptOutCheck, an end-to-end automated framework that detects inconsistencies between the actual data practices of online trackers and their policy statements regarding user opt-out choices. We have classified opt-out policies and created automatic NLP-based classifiers to extract the policies from trackers' opt-out web pages. OptOutCheck identifies opt-out buttons, detects opt-out cookies, and extracts the data flows sent to tracker servers. Finally, we have constructed a formal model to detect the inconsistencies. A large-scale study shows that trackers still continue the same data practices that contradict their stated opt-out policies even though these inconsistencies are violations of the trackers' own policies and may lose the users' trust in their services. OptOutCheck has laid a foundation for automatically detecting discrepancies between opt-out choices and actual data practices of online services. This is the extended version of the paper with the same title published at ACM CCS 2022.

ACKNOWLEDGMENTS

The work reported in this paper was supported in part by the US Army Research Office (ARO) under Grant No. W911NF-21-1-0057 and the US Office of Naval Research (ONR) under Grant No. N00014-22-1-2622.

REFERENCES

[1] Gunes Acar, Steven Englehardt, and Arvind Narayanan. 2020. No boundaries: data exfiltration by third parties embedded on web pages. *Proceedings on Privacy Enhancing Technologies (PETS)*, 2020, 4, 220–238.

[2] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. 2014. The web never forgets: persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, USA, 674–689. doi: 10.1145/2660267.2660347.

[3] Eugene Agichtein and Luis Gravano. 2000. Snowball: extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*. ACM, San Antonio, Texas, USA, 85–94. doi: 10.1145/336597.336644.

[4] Explosion AI. 2020. spaCy · industrial-strength natural language processing in python. Retrieved 01/08/2021 from https://spacy.io/.

[5] AllAboutCookies.org. 2021. What is an opt out cookie? - all about cookies. Retrieved 07/23/2021 from https://www.allaboutcookies.org/manage-cookies/opt-out-cookies.html.

[6] AllenAI. 2020. AllenNLP - Semantic Role Labeling. Retrieved 05/24/2020 from https://demo.allennlp.org/semantic-role-labeling.

[7] Digital Advertising Alliance. 2021. WebChoices tool. Retrieved 04/15/2021 from https://optout.aboutads.info/.

[8] Ryan Amos, Gunes Acar, Elena Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy policies over time: curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021*. ACM, New York, NY, USA, 2165–2176. doi: 10.1145/3442381.3450048.

[9] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *28th USENIX Security Symposium*. USENIX Association, Santa Clara, CA, 585–602.

[10] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions Speak Louder than Words: Entity-Sensitive Privacy Policy and Data Flow Analysis with POLICHECK. In *29th USENIX Security Symposium*, 18.

[11] Rebecca Balebako, Pedro G. Leon, Richard Shay, Blase Ur, Yang Wang, and Lorrie Faith Cranor. 2012. Measuring the effectiveness of privacy tools for limiting behavioral advertising. In *In Web 2.0 Workshop on Security and Privacy*.

[12] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, Florian Schaub, and Norman Sadeh. 2020. Finding a choice in a haystack: automatic extraction of opt-out statements from privacy policy text. In *Proceedings of The Web Conference 2020*. ACM, New York, NY, USA, 1943–1954. doi: 10.1145/3366423.3380262.

[13] Adam Barth. 2011. HTTP state management mechanism. Retrieved 12/31/2020 from https://tools.ietf.org/html/rfc6265.

[14] Berstend. 2021. Puppeteer-extra-plugin-stealth. Retrieved 06/30/2021 from https://www.npmjs.com/package/puppeteer-extra-plugin-stealth.

[15] Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021, 2.

[16] Duc Bui, Brian Tang, and Kang G. Shin. 2022. OptOutCheck datasets and extended paper version. https://github.com/ducalpha/optoutcheck_ccs22.

[17] Duc Bui, Brian Tang, and Kang G. Shin. 2022. OptOutCheck findings. https://rtcl.eecs.umich.edu/optoutcheck.

[18] Duc Bui, Yuan Yao, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Consistency analysis of data-usage purposes in mobile apps. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, Virtual Event, Republic of Korea. doi: 10.1145/3460120.3484536.

[19] CookiePro by OneTrust. 2020. What is a third-party cookie? CookiePro. Retrieved 03/15/2021 from https://www.cookiepro.com/knowledge/what-is-a-third-party-cookie/.

[20] Digital Advertising Alliance Of Canada. 2020. Online interest-based advertising FAQ. Retrieved 12/09/2020 from https://youradchoices.ca/en/faq.

[21] Quan Chen, Panagiotis Ilia, Michalis Polychronakis, and Alexandros Kapravelos. 2021. Cookie swap party: abusing first-party cookies for web tracking. In *Proceedings of the Web Conference 2021*. ACM, New York, NY, USA, 2117–2129. doi: 10.1145/3442381.3449837.

[22] Chris Coyier. 2021. Make entire div clickable. CSS-Tricks. Retrieved 03/01/2021 from https://css-tricks.com/snippets/jquery/make-entire-div-clickable/.

[23] Federal Trade Commission. 2011. FTC Puts an End to Tactics of Online Advertising Company That Deceived Consumers Who Wanted to "Opt Out" from Targeted Ads. Retrieved 12/07/2020 from https://www.ftc.gov/news-events/press-releases/2011/03/ftc-puts-end-tactics-online-advertising-company-deceived.

[24] Federal Trade Commission. 2012. Google Will Pay $22.5 Million to Settle FTC Charges it Misrepresented Privacy Assurances to Users of Apple's Safari Internet Browser. Retrieved 12/07/2020 from https://www.ftc.gov/news-events/press-releases/2012/08/google-will-pay-225-million-settle-ftc-charges-it-misrepresented.

[25] Federal Trade Commission. 2011. Online Advertiser Settles FTC Charges ScanScout Deceptively Used Flash Cookies to Track Consumers Online. Retrieved 12/07/2020 from https://www.ftc.gov/news-events/press-releases/2011/11/online-advertiser-settles-ftc-charges-scanscout-deceptively-used.

[26] Cookiebot. 2022. Google ending third-party cookies in Chrome. Retrieved 08/08/2022 from https://www.cookiebot.com/en/google-third-party-cookies/.

[27] CookiePro. 2021. What is an opt-out cookie? CookiePro. Retrieved 07/23/2021 from https://www.cookiepro.com/knowledge/what-is-an-opt-out-cookie/.

[28] Lorrie Faith Cranor, Candice Hoke, Pedro Giovanni Leon, and Alyssa Au. 2015. Are they worth reading? an in-depth analysis of online trackers' privacy policies. *I/S: a journal of law and policy for the information society*.

[29] Crownpeak Technology, Inc. 2020. Global consent preferences. Global Consent Preferences. Retrieved 12/02/2020 from https://www.evidon.com/resources/global-opt-out/.

[30] Chrome Developers. 2021. Chrome.cookies API. Retrieved 01/09/2021 from https://developer.chrome.com/docs/extensions/reference/cookies/.

[31] Disconnect. 2021. Disconnectme/disconnect-tracking-protection. Retrieved 06/27/2021 from https://github.com/disconnectme/disconnect-tracking-protection.

[32] Disconnect. 2021. Tracking protection lists. Retrieved 06/26/2021 from https://disconnect.me/trackerprotection.

[33] DuckDuckGo. 2022. DuckDuckGo tracker radar. Retrieved 04/23/2022 from https://github.com/duckduckgo/tracker-radar.

[34] EasyList. 2021. EasyList Filter. Retrieved 06/26/2021 from https://easylist.to/.

[35] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: a 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS'16: 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, Vienna Austria, 1388–1401. doi: 10.1145/2976749.2978313.

[36] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. 2015. Cookies That Give You Away: The Surveillance Implications of Web Tracking. In *Proceedings of the 24th International Conference on World Wide Web*. IW3C2, Republic and Canton of Geneva, CHE, 289–299. doi: 10.1145/2736277.2741679.

[37] IAB Europe. 2020. Transparency and consent framework. GitHub. Retrieved 03/14/2021 from https://github.com/InteractiveAdvertisingBureau/GDPR-Transparency-and-Consent-Framework.

[38] Evidon. 2021. Your opt-out choices. Evidon. Retrieved 07/23/2021 from https://www.evidon.com/opting-out/.

[39] Imane Fouad, Cristiana Santos, Arnaud Legout, and Nataliia Bielova. 2022. My Cookie is a phoenix: detection, measurement, and lawfulness of cookie respawning with browser fingerprinting. In *PETS 2022 - 22nd Privacy Enhancing Technologies Symposium*. Sydney, Australia.

[40] Mozilla Foundation. 2020. Public suffix list. Retrieved 03/15/2021 from https://publicsuffix.org/.

[41] Gertjan Franken, Tom Van Goethem, and Wouter Joosen. 2018. Who left open the cookie jar? a comprehensive evaluation of third-party cookie policies. In 27th USENIX Security Symposium, 151–168.

[42] Adtriba GmbH. 2021. Adtribas cookies. Retrieved 07/20/2021 from http://help.adtriba.com/en/articles/3772978-adtribas-cookies.

[43] Cliqz GmbH. 2017. Tracker categories. Retrieved 06/27/2021 from https://whotracks.me/blog/tracker_categories.html.

[44] Cliqz GmbH. 2021. WhoTracks.me - bringing transparency to online tracking. Retrieved 03/15/2021 from https://whotracks.me.

[45] R. Gonzalez, L. Jiang, M. Ahmed, M. Marciel, R. Cuevas, H. Metwalley, and S. Niccolini. 2017. The cookie recipe: untangling the use of cookies in the wild. In *2017 Network Traffic Measurement and Analysis Conference (TMA)*, 1–9. doi: 10.23919/TMA.2017.8002896.

[46] Google. 2015. Keep my opt-outs. Retrieved 12/23/2020 from https://github.com/google/chrome-opt-out-extension.

[47] Google Chrome DevTools Team. 2020. Puppeteer Tools for Web Developers. Retrieved 02/05/2020 from https://pptr.dev/.

[48] Google Inc. 2021. Custom search JSON API. Retrieved 07/09/2021 from https://developers.google.com/custom-search/v1/reference/rest/v1/cse/list.

[49] Google Inc. 2021. Programmable search engine. Retrieved 07/08/2021 from https://programmablesearchengine.google.com/.

[50] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2020. "It's a scavenger hunt": Usability of Websites' Opt-Out and Data Deletion Choices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. doi: 10.1145/3313831.3376511.

[51] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2019. An empirical analysis of data deletion and opt-out choices on 150 websites. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security*. USENIX Association, USA, 387–406.

[52] Jeff Huang, Oren Etzioni, Luke Zettlemoyer, Kevin Clark, and Christian Lee. 2012. RevMiner: an extractive interface for navigating reviews on a smartphone. In *Proceedings of ACM symposium on User interface software and technology*. ACM, New York, NY, USA, 3–12. doi: 10.1145/2380116.2380120.

[53] Jeff Huang, Patrick O'Neil Meredith, and Grigore Rosu. 2014. Maximal sound predictive race detection with control flow abstraction. *ACM SIGPLAN Notices*, 49, 6, 337–348. doi: 10.1145/2666356.2594315.

[54] Ilya Sergey. 2019. What does it mean for a program analysis to be sound? SIGPLAN Blog. Retrieved 10/03/2020 from https://blog.sigplan.org/2019/08/07/what-does-it-mean-for-a-program-analysis-to-be-sound/.

[55] Google Inc. 2021. Robots.txt introduction & guide | google search central. Google Developers. Retrieved 07/09/2021 from https://developers.google.com/search/docs/advanced/robots/intro.

[56] Moz Inc. 2021. Page speed best practices. Moz. Retrieved 03/14/2021 from https://moz.com/learn/seo/page-speed.

[57] Network Advertising Initiative. 2020. FAQ | NAI: network advertising initiative. Retrieved 03/14/2021 from https://www.networkadvertising.org/faq.

[58] Network Advertising Initiative. 2021. NAI consumer opt out. Retrieved 04/15/2021 from https://optout.networkadvertising.org/.

[59] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. 2018. Tracing cross border web tracking. In *Proceedings of the Internet Measurement Conference 2018*. ACM, New York, NY, USA, 329–342. doi: 10.1145/3278532.3278561.

[60] Haojian Jin, Minyi Liu, Kevan Dodhia, Yuanchun Li, Gaurav Srivastava, Matthew Fredrikson, Yuvraj Agarwal, and Jason I. Hong. 2018. Why Are They Collecting My Data?: Inferring the Purposes of Network Traffic in Mobile Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2, 4, 173. doi: 10.1145/3287051.

[61] Haojian Jin, Tetsuya Sakai, and Koji Yatani. 2014. ReviewCollage: a mobile interface for direct comparison using online reviews. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. ACM, New York, NY, USA, 349–358. doi: 10.1145/2628363.2628373.

[62] Garrett A. Johnson, Scott K. Shriver, and Shaoyin Du. 2020. Consumer Privacy Choice in Online Advertising: Who Opts Out and at What Cost to Industry? *Marketing Science*, 39, 1, 33–51. doi: 10.1287/mksc.2019.1198.

[63] Jordan Jueckstock, Shaown Sarker, Peter Snyder, Aidan Beggs, Panagiotis Papadopoulos, Matteo Varvello, Ben Livshits, and Alexandros Kapravelos. 2021. Towards Realistic and Reproducible Web Crawl Measurements. In *Proceedings of the The Web Conference*.

[64] Daniel Jurafsky and James H. Martin. 2019. *Speech and Language Processing, 3rd edition*. (Third Edition draft edition).

[65] Arjaldo Karaj, Sam Macbeth, Rémi Berson, and Josep M. Pujol. 2019. WhoTracks.me: shedding light on the opaque world of online tracking. *arXiv:1804.08959 [cs]*. arXiv: 1804.08959.

[66] Saranga Komanduri, Richard Shay, Greg Norcie, Blase Ur, and Lorrie Faith Cranor. 2011. AdChoices? Compliance with Online Behavioral Advertising Notice and Choice Requirements. *I/S: A Journal of Law and Policy for the Information Society*, 7, 603.

[67] John Kurkowski. 2020. Tldextract. Retrieved 03/15/2021 from https://github.com/john-kurkowski/tldextract.

[68] Daniil Larionov, Artem Shelmanov, Elena Chistova, and Ivan Smirnov. 2019. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. INCOMA Ltd., Varna, Bulgaria, 619–628. doi: 10.26615/978-954-452-056-4_073.

[69] Richard Lawler. 2022. Google delays blocking third-party cookies again, now targeting late 2024. Retrieved 08/08/2022 from https://www.theverge.com/2022/7/27/23280905/google-chrome-cookies-privacy-sandbox-advertising.

[70] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: a research-oriented top sites ranking hardened against manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*. doi: 10.14722/ndss.2019.23386.

[71] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. 2016. Internet jones and the raiders of the lost trackers: an archaeological study of web tracking from 1996 to 2016. In *25th USENIX Security Symposium*.

[72] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

[73] Will Bontrager Software LLC. 2021. Linking without an 'a' tag. Retrieved 03/01/2021 from https://willmaster.com/library/web-development/linking-without-an-a-tag.php.

[74] SimilarWeb LTD. 2020. Top websites in united states. Retrieved 11/28/2020 from https://www.similarweb.com/top-websites/united-states/.

[75] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[76] C. Matte, N. Bielova, and C. Santos. 2020. 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*, 791–809. doi: 10.1109/SP40000.2020.00076.

[77] Hassan Metwalley, Stefano Traverso, and Marco Mellia. 2015. Unsupervised detection of web trackers. In *2015 IEEE Global Communications Conference*, 1–6. doi: 10.1109/GLOCOM.2015.7417499.

[78] Microsoft. 2020. Microsoft/playwright-python. Retrieved 12/18/2020 from https://github.com/microsoft/playwright-python.

[79] Microsoft. 2021. Tracking prevention. Retrieved 06/26/2021 from https://docs.microsoft.com/en-us/microsoft-edge/web-platform/tracking-prevention.

[80] Moz Inc. 2021. URL structure. Moz. Retrieved 07/09/2021 from https://moz.com/learn/seo/url.

[81] Mozilla. 2021. Enhanced tracking protection in firefox for desktop. Retrieved 06/26/2021 from https://support.mozilla.org/en-US/kb/enhanced-tracking-protection-firefox-desktop#w_what-enhanced-tracking-protection-blocks.

[82] Mozilla. 2021. Navigation timing API - web APIs | MDN. Retrieved 03/14/2021 from https://developer.mozilla.org/en-US/docs/Web/API/Navigation_timing_API.

[83] Mozilla. 2021. Node.textContent - web APIs | MDN. Retrieved 07/09/2021 from https://developer.mozilla.org/en-US/docs/Web/API/Node/textContent.

[84] Mozilla. 2021. Window: load event - web APIs | MDN. Retrieved 03/14/2021 from https://developer.mozilla.org/en-US/docs/Web/API/Window/load_event.

[85] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In EMNLP. Association for Computational Linguistics, 2774–2779.

[86] Niclas. 2020. Clefspeare13/pornhosts. Retrieved 12/13/2020 from https://github.com/Clefspeare13/pornhosts.

[87] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. 2019. Cookie synchronization: everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference*. ACM, New York, NY, USA, 1432–1442. doi: 10.1145/3308558.3313542.

[88] Paul E. Black. 2004. Ratcliff/obershelp pattern recognition. Retrieved 01/02/2021 from https://xlinux.nist.gov/dads/HTML/ratcliffObershelp.html.

[89] Python-Markdown. 2021. Markdown. Retrieved 07/01/2021 from https://github.com/Python-Markdown/markdown.

[90] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0 LDC2013T19. (2013).

[91] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and Defending Against Third-Party Tracking on the Web. In 155–168.

[92] T. Sakamoto and M. Matsunaga. 2019. After GDPR, Still Tracking or Not? Understanding Opt-Out States for Online Behavioral Advertising. In *2019 IEEE Security and Privacy Workshops (SPW)*, 92–99. doi: 10.1109/SPW.2019.00027.

[93] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. 2019. 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*. ACM, New York, NY, USA, 340–351. doi: 10.1145/3321705.3329806.

[94] Alireza Savand. 2021. Html2text. Retrieved 07/01/2021 from https://github.com/Alir3z4/html2text.

[95] SEOPressor. 2019. Does URL structure affect SEO? Retrieved 07/09/2021 from http://seopressor.com/blog/url-structure-affect-seo/.

[96] Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv:1904.05255 [cs]*.

[97] Anastasia Shuba and Athina Markopoulou. 2020. NoMoATS: Towards Automatic Detection of Mobile Tracking. *Proceedings on Privacy Enhancing Technologies*, 2020, 2, 45–66. doi: 10.2478/popets-2020-0017.

[98] Yannis Smaragdakis, Jacob Evans, Caitlin Sadowski, Jaeheon Yi, and Cormac Flanagan. 2012. Sound predictive race detection in polynomial time. In *Proceedings of the 39th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. ACM, New York, NY, USA, 387–400. doi: 10.1145/2103656.2103702.

[99] Sovrn. 2018. Sovrn acquires VigLink to expand publisher services. Sovrn, Inc. Retrieved 05/01/2022 from https://www.sovrn.com/blog/sovrn-acquires-viglink/.

[100] Statista. 2022. Internet users in the world 2022. Statista. Retrieved 04/28/2022 from https://statista.com/statistics/617136/digital-population-worldwide/.

[101] Taboola. 2022. Taboola Access Request. Retrieved 08/02/2022 from https://web.archive.org/web/20220710192140/https://accessrequest.taboola.com/access.

[102] Tim Nolet. 2020. Navigating & waiting. Retrieved 03/14/2021 from https://theheadless.dev/posts/basics-navigation/.

[103] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Beyond the front page:measuring third party dynamics in the field. In *Proceedings of The Web Conference 2020*. ACM, New York, NY, USA, 1275–1286. DOI: 10.1145/3366423.3380203.

[104] Pelayo Vallina, Alvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. 2019. Tales from the porn: a comprehensive privacy analysis of the web porn ecosystem. In *Proceedings of the Internet Measurement Conference*. ACM, New York, NY, USA, 245–258. DOI: 10.1145/3355369.3355583.

[105] W3C. 2012. Navigation Timing. Retrieved 03/14/2021 from https://www.w3.org/TR/navigation-timing/.

[106] Zhiju Yang and Chuan Yue. 2020. A comparative measurement study of web tracking on mobile and desktop environments. *Proceedings on Privacy Enhancing Technologies*, 2020, 2, 24–44. DOI: 10.2478/popets-2020-0016.

A WEB CRAWLER TIMEOUTS

To avoid missing cookies due to too short experiment time on web pages with dynamically loaded resources, we design the following timeout values. For each web page, the crawler navigates to a URL and waits until there is no active connection (i.e., no active resource downloads) for 500ms [102]. This timeout is effective for dynamically loaded resources (e.g., advertisements in iframes) and typically much later than the completion of the page load event (i.e., when the entire page, including all dependent resources such as stylesheets and images, has been loaded [84]). If the page has always-active connections, the crawler stops loading after a final timeout of 30 seconds. This timeout is sufficient to avoid missing resources because the crawler runs on a server with a fast network speed while the top websites are commonly designed to reduce page load time for a large number of visitors [56]. In particular, the page load time (i.e., duration from Navigation Start to the Load Event End [82, 105]) of websites in the top 5k domains is 2.47 (2.28 SD) seconds. In the remainder of this paper, unless stated otherwise, we use these timeout heuristics in our experiments.

B AUTOMATICALLY CLICKING A BUTTON

To click on an opt-out button, we use two complement methods: the *click()* function provided by the automation tool and the JavaScript *click()* API. The automation tool's version scrolls the button to the viewport and checks the visibility of the button before issuing a click event to the element. However, this method fails when the button is hidden by a popup menu such as an email-subscription banner when the page is first loaded. The JavaScript version bypasses the checks and can activate the opt-out button in this case. However, in some cases where the click is intercepted by an outer HTML element, the JavaScript version can still fail. When both clicking methods fail, `OptOutCheck` concludes the button is not clickable.

C OPT-OUT POLICY CORPUS

C.1 Cookie Domain Selection

The cookie domains in the opt-out policy corpus are selected as follows. The data collection of the top 5k websites crawled 2.2M cookies from 38.8k web pages on 4,437 websites. Of these, 1.6M third-party cookies from 2,364 unique cookie domains were collected. We removed the cookie domains present on less than 100 websites to reduce the noise from the first-party cookies of publishers placed on their own domains. We also excluded ad platforms that do not provide any English version of their websites and policies. After this removal, 180 cookie domains remained. We selected the top 100 cookie domains and randomly selected additional 20 domains from the rest of the list to cover both ends of the spectrum of ad platforms in terms of popularity.

C.2 Opt-out Button Identification

We searched opt-out settings of an ad platform as follows. We first looked for the privacy policies of the advertiser. Starting from the home page, we find keywords describing privacy policies, such as the combination of *{privacy, cookie}* and *{policy, notice, statement}*.

In the policies found in the previous step, we searched for the keywords "opt-out" and "opt out", and read through the surrounding text to check whether it mentioned the opt-out choices for users from OBA, data collection or tracking of the ad platforms, rather than on its website. It is not straightforward to identify the direct opt-out settings because they commonly provide separate privacy policies for their websites themselves (now they are the first party) and for the websites where they place their cookies as a third party.

If no opt-out option was found in the policies, we searched for the opt-out page from the home page. We looked for opt-out keywords such as "opt-out", "opt out", "consumer choice", "interest based ads", and "ad choice". These steps are necessary because the settings are not always in the privacy policy but could be located on a dedicated ad choice page. The websites sometimes placed a direct link with the text "opt out" on their home pages. We also looked for opt-out choices in the "What choices do I have?" and the like.

If the opt-out button could not be found in the privacy policies and home page, we looked for "opt out opt-out site:*tracker_domain*" on a search engine to find the opt-out page on the tracker's website. This process was repeated until an opt-out setting was found or the advertiser was determined not to provide such a setting.

C.3 Opt-out Policy Classifier Performance

Table 7 shows the performance of the opt-out policy classifiers on the opt-out policy corpus.

Opt-out Policy	Precision	Recall	F1	Support
No-tracking	88.24	83.33	85.71	18
No-data-collection	90.48	82.61	86.36	23
Average	89.36	82.97	86.04	20

Table 7: Performance of the opt-out policy classifiers on the opt-out policy corpus.

D PROOF OF THEOREM 8.4

Proof of Theorem 8.4 is as follows.

PROOF. The collection of unique IDs for tracking purposes constitutes a data flow $f = (r, unique\_id, tracking)$ where $r \equiv_{\delta}$ "first party", according to Definition 7.1.

Informally, the data flow is inconsistent with *No-tracking* policy due to their contradictory data-usage purposes. From the definitions of policy classes in Table 1, there exists a policy statement $((r, collect, data), (data, not\_for, tracking))$. Since a unique ID is a type of user data, i.e., $unique\_id \sqsubseteq data$, the statement is flow-relevant to f but $k_t = not\_for$. So the policy is inconsistent with the flow by Definition 8.3.

Similarly, the data flow is inconsistent with *No-data-collection* policy because it collects a data type that is not allowed by the policy. By definition of the policy classes, there exists a statement $((r, not\_collect, data), None)$ in the *No-data-collection* opt-out class. Because a unique ID is a type of user data, i.e., $unique\_id \sqsubseteq data$, the policy is flow-relevant to f but $c_t = not\_collect$, and thus, is inconsistent with the flow by Definition 8.3. □

E TRACKER DATABASES

The selected tracker databases contain 229–3,194 trackers as shown in Table 8. The WhoTracksMe database contains trackers from usage data collected via the Ghostery extension's users from May 2017 to March 2022. The Disconnect database is created and updated by using manual reviews of trackers' scripts/privacy policies and error reports from the companies labeled as trackers [31]. The Evidon Global Opt-out tool contained 796 trackers at the time of this writing. Finally, we extracted privacy-policy URLs of 229 trackers from the 2022 March crawl of the DuckDuckGo Tracker Radar database.

Tracker Database	# Trackers
WhoTracksMe	3,194
Disconnect	1,393
Evidon	796
DuckDuckGo	229
Merge	4,021

Table 8: Sizes of the tracker databases.

F "OTHER" PRIVACY STATEMENTS

Table 9 lists the most common categories of "Other"-labeled privacy policy statements with a frequency $\geq$ 30 sentences.

High-level	Low-level	# Sents.
Opt-out Explanation	Opt-Out Mechanism	66
	Step-by-Step Guide	58
	Technical Requirement	55
	Status of Opt-Out	37
	Total	216
Data collection	Data Usage Purposes	73
	Collected Data	71
	Cookie Usage	49
	Total	193
Legality	Legal Statement	154
Scope	Opt-Out Scope	49
	Opt-Out Is Device-Specific	48
	Total	97
General Description	Product Description	88
Other Opt-out	Disable Browser Cookies	32

Table 9: The most common categories of "Other" statements with $\geq$ 30 sentences. The first two columns are the high- and low-level categories.

G DETECTED INCONSISTENT TRACKERS

The extracted inconsistent data flows, opt-out policies, opt-out cookies and opt-out domains of the detected inconsistent trackers are listed in Tables 10, 11, 12, and 13, respectively.

H EMAIL TEMPLATE

Subject: Issues with <TRACKER_NAME> Opt-Out
Dear <TRACKER_NAME> privacy team,
We are security researchers from [Redacted for anonymous review].
We have developed a tool to verify the opt-out choices of online trackers and found that your opt-out choices might not work as intended/stated as described below.
We expect the termination of tracking after opting out because your privacy policy states the following: <OPT-OUT_POLICY_SENTENCE(S)>. However, your tracking cookie <COOKIE_NAME> (domain <COOKIE_DOMAIN>) with a <DATA_TYPE> is still retained after the opt-out.
On June 18, 2022, we performed the following steps to observe the inconsistent behavior:
- Step 1: Using a clean instance of Google Chrome, we browsed popular publisher sites (such as <WEBSITE_1>, <WEBSITE_2>, and <WEBSITE_3>) where your tracking cookies were placed.
- Step 2: We then clicked the opt-out button on your opt-out page at <TRACKER_OPT-OUT_PAGE>. An opt-out cookie <OPT-OUT_COOKIE> (domain <OPT-OUT_COOKIE_DOMAIN>) was created.
- Step 3: However, the publisher sites still retained and used the tracking cookie(s) with unique values.
Don't these steps demonstrate an inconsistency between your opt-out policy and the actual tracking behavior?
We would appreciate it if you could provide us with more information to clarify this issue, preferably by June 28, 2022.
Best regards,
<EMAIL_SIGNATURE>

Figure 6: Email template for a *No-tracking* opt-out policy.

Tracker	Opt-out Policies	Data Type	Cookie Name	Cookie Domain
adtriba.com	*No-tracking*	Unique ID	atbgdid	.adtriba.com
criteo.com	*No-data-collection*	Unique ID	uid	.criteo.com
		Unique ID	uid	.storetail.io
deepintent.com	*No-data-collection*	Unique ID	CDIUSER	.deepintent.com
dianomi.com	*No-tracking*	Unique ID	session[1]	.dianomi.com
dynad.net	*No-data-collection*, *No-tracking*	Unique ID	uid	.dynad.net
liveintent.com	*No-data-collection*	Unique ID	lidid	.liadm.com
onaudience.com	*No-data-collection*	Unique ID	cookie	.onaudience.com
reachlocal.com	*No-tracking*	Unique ID	visitor_id	04be4b16-c90f-4f1c-89a3-f7f1c516e394.rlets.com
		Unique ID	visitor_id	789a4467-dcc5-452e-9434-e15256aed01b.rlets.com
sovrn.com	*No-data-collection*, *No-tracking*	Unique ID	vglnk.Agent.p	.viglink.com[2]
taboola.com	*No-tracking*	Unique ID	t_gid	.taboola.com
underdogmedia.com	*No-data-collection*	Location	geode	.udmserve.net
		Unique ID	apnid	.udmserve.net
		Unique ID	pmid	.udmserve.net
		Unique ID	sncr	.udmserve.net

Table 10: Detected inconsistencies and data flows. [1] Despite the name, this is a persistent cookie. [2] Sovrn acquired VigLink and owned viglink.com [99].

Tracker	Sentence	Opt-out Policy
adtriba.com	It is at all times possible to object to the data collection through this third party tracking and storage with effect for the future (Opt - Out) .	*No-tracking*
	In order to be excluded from Adtriba third party tracking , you can click the following button .	*No-tracking*
	[Opt-out Button] Opt - Out from Adtriba tracking	*No-tracking*
criteo.com	[Opt-out Button] Disable Criteo services	*No-data-collection*
deepintent.com	This page is intended to help you opt out of the use of cookies , and other data points .	*No-data-collection*
dianomi.com	Opted Out : you have opted out of tracking (behavioural targeting) , Dianomi will no longer serve you with personalized content recommendations based on your Internet history .	*No-tracking*
dynad.net	Opting out of DynAd services through this link from the User 's browser inserts a cookie on the User 's browser and clear all LSOs stored data by DynAd Service .	*No-data-collection*
	This means that we will not track an opted out User 's behavior or display customized ads to the User .	*No-tracking*
liveintent.com	You can opt - out of the cookie - based portion of the LiveIntent Advertising Program by clicking [Opt-out Button] here .	*No-data-collection*
onaudience.com	An “ opt - out cookie ” will be installed in your browser and block the placement of cookies from OnAudience .	*No-data-collection*
reachlocal.com	If you would like to opt out of tracking provided by the ReachLocal Tracking Code , click the button below .	*No-tracking*
sovrn.com	Opting out of Sovrn //Commerce cookies means that Sovrn //Commerce will stop placing cookies on your device when you browse Sovrn //Commerce enabled websites and/or links .	*No-data-collection*
	To honor your opt - out choice , a cookie is required on your device so that we know not to track your activity .	*No-tracking*
	[Opt-out Button] Disable Tracking	*No-tracking*
taboola.com	* Opted Out : You have opted out of tracking , Taboola will no longer serve you with personalized content recommendations based on your Internet use history .	*No-tracking*
underdogmedia.com	Your current status is to opt - out for Underdog Media hosted 3rd Party Cookies .	*No-data-collection*

Table 11: Opt-out policies of the detected inconsistent trackers. *[Opt-out-Button]* indicates the occurrence of an opt-out button.

	Opt-out Cookies		
Tracker	Domain	Value	Names
adtriba.com	.adtriba.com	1	atboptout
criteo.com	.courses-en-ligne.carrefour.fr	1	STO_carrefourv2_optout, STO_carrefour_one_optout
	.criteo.com	1	optout
	.fnac.com	1	STO_fnac_optout
	.hlserve.com	1	oo
	.laredoute.fr	1	STO_laredoute_optout
	.storetail.io	1	STO_carrefour_espagne_v2_optout, STO_clarel_optout, STO_alcampo_optout, STO_dia_es_optout, STO_metro_en_optout, STO_metro_fr_optout, STO_phone_house_es_optout, STO_carrefour_espagne_optout, STO_fnac_espagne_optout, STO_ulabox_optout, STO_tudespensa_optout, STO_planetahuerto_optout, STO_pccomponentes_optout, STO_worten_es_optout, STO_worten_pt_optout, STO_primor_optout, STO_costco_optout, STO_auchandrive_v2_optout, STO_cora_optout, STO_auchan_optout, STO_darty_optout, STO_croquetteland_optout, STO_delhaize_fr_optout, STO_delhaize_nl_optout, STO_fnac_belgique_fr_optout, STO_fnacspectacles_optout, STO_fnac_portugal_optout, STO_intermarche_optout, STO_fnac_belgique_nl_optout, STO_leclercculture_optout, STO_sephora_optout, STO_leclerc_optout, STO_leclercpara_optout, STO_sephora_mobile_optout, STO_leclercportal_optout, STO_micromania_optout, STO_rewe_optout, STO_leroymerlin_optout, STO_kingjouet_optout, STO_auchan_v2_optout
deepintent.com	.deepintent.com	true	optout
dianomi.com	.dianomi.com	1	dnt
dynad.net	.dynad.net	1	optout
liveintent.com	d.liadm.com	opt-out	tuuid
onaudience.com	www.onaudience.com	1	opt-out
reachlocal.com	.rlets.com	1	RlocalOptOut
sovrn.com	.sovrn.co	1	vglnk.OptOut.p
	.viglink.com	1	vglnk.OptOut.p
taboola.com	.taboola.com	1	DNT
underdogmedia.com	.udmserve.net	Thank_You	optout

Table 12: Opt-out cookies of the detected inconsistent trackers. These cookies are grouped by their domains and values. All cookies have a "/" path.

Tracker	Opt-out Domains	Opt-out Page URL
adtriba.com	.adtriba.com	https://privacy.adtriba.com/
criteo.com	.carrefour.fr, .criteo.com, .fnac.com, .hlserve.com, .laredoute.fr, .storetail.io	https://www.criteo.com/privacy/disable-criteo-services-on-internet-browsers/
deepintent.com	.deepintent.com	https://option.deepintent.com/optout
dianomi.com	.dianomi.com	https://www.dianomi.com/legal/privacy.epl
dynad.net	.dynad.net	https://www.dynad.net/en/privacy-and-terms.html
liveintent.com	.liadm.com	https://www.liveintent.com/ad-choices/
onaudience.com	.onaudience.com	https://www.onaudience.com/opt-out
reachlocal.com	.rlets.com	https://www.reachlocal.com/us/en/legal/trackingopt-out
sovrn.com	.sovrn.co, .viglink.com	https://www.sovrn.com/legal/privacy-center/
taboola.com	.taboola.com	https://www.taboola.com/policies/privacy-policy
underdogmedia.com	.udmserve.net	https://underdogmedia.com/optout/

Table 13: Opt-out domains and web page URLs of the detected inconsistent trackers.