

# An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies

Timothy Libert

Reuters Institute for the Study of Journalism  
Department of Computer Science  
University of Oxford  
Oxford, United Kingdom  
public@timlibert.me

## ABSTRACT

A dominant regulatory model for web privacy is “notice and choice”. In this model, users are notified of data collection and provided with options to control it. To examine the efficacy of this approach, this study presents the first large-scale audit of disclosure of third-party data collection in website privacy policies. Data flows on one million websites are analyzed and over 200,000 websites’ privacy policies are audited to determine if users are notified of the names of the companies which collect their data. Policies from 25 prominent third-party data collectors are also examined to provide deeper insights into the totality of the policy environment. Policies are additionally audited to determine if the choice expressed by the “Do Not Track” browser setting is respected.

Third-party data collection is wide-spread, but fewer than 15% of attributed data flows are disclosed. The third-parties most likely to be disclosed are those with consumer services users may be aware of, those without consumer services are less likely to be mentioned. Policies are difficult to understand and the average time requirement to read both a given site’s policy and the associated third-party policies exceeds 84 minutes. Only 7% of first-party site policies mention the Do Not Track signal, and the majority of such mentions are to specify that the signal is ignored. Among third-party policies examined, none offer unqualified support for the Do Not Track signal. Findings indicate that current implementations of “notice and choice” fail to provide notice or respect choice.

## CCS CONCEPTS

• **Security and privacy** → **Human and societal aspects of security and privacy; Usability in security and privacy;**

## KEYWORDS

Web Privacy, Web Security, Internet Policy, Internet Regulation

### ACM Reference Format:

Timothy Libert. 2018. An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3178876.3186087>

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

*WWW 2018, April 23–27, 2018, Lyon, France*

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5639-8/18/04.

<https://doi.org/10.1145/3178876.3186087>

## 1 INTRODUCTION

Although many users may not be aware, web pages are not unitary objects downloaded directly from the party listed in a browser’s address bar. Rather, most web pages are a collection of media elements which are either downloaded from the first-party a user is aware of, or from third-parties a user may not know of. When a page includes third-party content, HTTP “Referer” headers convey the address of the page the user is currently visiting to third-parties. While users may be happy to see third-party content in a web page, they may not be happy that such content allows third-parties to create records of their browsing behaviors. The process of using third-party HTTP Referer headers to observe users’ web browsing is often referred to as “web tracking”.

Research has demonstrated that web pages often expose a user’s browsing history to numerous third-parties in tandem [12, 22–24]. Quite often these parties collect data on users’ behavior so that they may be shown advertisements which are tailored to their interests. The benefit of this system is that users may enjoy learning about products and services relevant to their lives, website operators are able to make the most efficient use of limited screen space, and vendors are able to directly reach potential customers. However, these commercial imperatives come at a cost to personal privacy and are poorly regulated.

Despite the fact that half of the top ten companies in the world ranked by market value conduct web tracking, there is very little formal oversight of the practice.<sup>1</sup> One reason for this is that the United States lacks a top-level data protection authority.<sup>2</sup> While nations in the European Union have designated data protection authorities, researchers have found that third-parties routinely ignore regulations designed to police the use of third-party tracking cookies and deem the approach a “failure” [39]. The data collection industry claims that formal oversight is not needed due to the adherence to a “self-regulatory” framework called “notice and choice”.

Under the notice and choice framework, users are theoretically notified that data collection is taking place and given options to control the practice (often called an “opt-out”). According to industry group Network Advertising Initiative, member companies follow a code which “requires notice and choice with respect to Interest-Based Advertising, limits the types of data that member

<sup>1</sup>According to Statista, Alphabet (Google), Microsoft, Amazon, Facebook, and Tencent are all in the top ten: <https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-value/>

<sup>2</sup>Although the U.S. Federal Trade Commission has been involved in online privacy for years, the primary remit of the agency is not data protection.

companies can use for advertising purposes, and imposes a host of substantive restrictions on member companies' collection, use, and transfer of data used for Interest-Based Advertising" [31]. Despite the fact that targeted advertising now supports the vast majority of online publications, social media sites, and search engines, there has been very little auditing to determine if self-regulatory frameworks are in fact being followed. A 2011 evaluation of compliance with industry-defined guidelines for notice and choice conducted by Komanduri et al is one of the most notable studies of the topic [20].

Although a large volume of academic literature has identified the parties which collect data on websites and deficiencies in the nature of online privacy policies, there has been virtually no attempt to determine if the parties that collect data on a given site are disclosed in the policy of that site. This study presents the first attempt at auditing disclosure of third-party data flows in website privacy policies and a new software tool, *policyxray*, is presented. *policyxray* facilitates large-scale auditing of privacy policies and has been used to determine if policies for 207,000 websites accurately disclose the third-parties which collect user data.

Privacy policies are also analyzed to determine if the text is easy to understand, how long the text would take to read, and if the "Do Not Track" choice mechanism is respected. Network traffic is inspected to determine if transport encryption is used. Finally, rather than treating third-parties as an undifferentiated whole, the policies and practices of 25 prominent data collectors are examined in order to reveal variations in practices.

Third-party data collection is wide-spread, but fewer than 15% of attributed data flows are disclosed. The third-parties most likely to be disclosed are those with consumer services users may be aware of, those without consumer services are mentioned in less than 1% of instances. Policies are difficult to understand and the average time requirement to read both a given site's policy and the associated third-party policies exceeds 84 minutes. Only 7% of first-party site policies mention the Do Not Track signal, and the majority of such mentions are to specify that the signal is ignored. Among third-party policies examined, none offer unqualified support for the Do Not Track signal. Findings indicate that current implementations of notice and choice fail to provide notice or respect choice.

## 2 RESEARCH QUESTIONS: NOTICE, CHOICE, AND SECURITY

The overarching purpose of this study is to evaluate the efficacy of the notice and choice policy regime. However, there are no commonly agreed upon definitions of what constitutes sufficient notice or choice on the web. In the United States this may be partially attributed to the fact that the Federal Trade Commission's guidance on the topic has been "consistently inconsistent" [15]. In the European Union, the *ePrivacy Directive* (sometimes referred to as the "cookie law") is being replaced by the *ePrivacy Regulation* in 2018 and there remains substantial uncertainty as to what changes will arise from the transition.

One potential metric for notice is the industry-favored approach of the "AdChoices" icon, a small blue arrow which sits in the corner of advertisements. When clicked upon, the icon will take a user to information about the party responsible for the ad. However,

not all third-parties show ads, thus AdChoices cannot offer full disclosure of all parties collecting data. Furthermore, researchers have found that "the purpose of these icons, to provide information to consumers, eluded participants, even when the icons were shown in context on an advertisement" [44].

It is possible that mentions of sharing data with undefined "third-parties" may be viewed as providing notice in the context of a privacy policy. However, given that users are subject to *both* first- and third-party privacy policies, this type of *ambiguous notice* does not provide users with a means to evaluate all policies to which they are subject. Indeed, different parties have different policies, and users must know the names of specific data collectors to exercise meaningful choice. Therefore, for the purposes of this study, merely mentioning ambiguous "third-party" data sharing does not qualify as meaningful notice.

In absence of agreed upon guidelines, purposefully limited questions have been asked in order to determine the degree to which users receive notice and are able to convey choices. Three decisions have been made to simplify the scope of the evaluation. First, in place of evaluating inconsistent icons and modal dialogues, this study evaluates human-readable privacy policies as the vehicle for notice. Second, in order to establish a benchmark for choice, mention of, and respect for, the "Do Not Track" (DNT) browser signal is evaluated. While the online advertising industry has advocated for many different forms of choice ranging from setting "opt-out" cookies to instructing users to disable third-party cookies,<sup>3</sup> DNT is the only signal common to all major browsers and its development was encouraged by the U.S. Federal Trade Commission [43].<sup>4</sup> Finally, in regards to the security of data transmission, the use of Secure Sockets Layer (SSL) transport encryption is measured.

Based on the above scoping decisions, the research questions being asked regarding **Notice** are as follows:

- Who are the third-parties which collect user data on websites, and do they have consumer services users may already be aware of?
- If users read a privacy policy from a given website will they learn of the specific third-parties which receive their data?
- How time-consuming and difficult is it to understand website privacy policies?
- How time-consuming and difficult is it to understand third-party privacy policies?

Questions regarding **Choice** are as follows:

- Do website privacy policies mention and respect Do Not Track signals?

<sup>3</sup>There is significant variability in how users may "opt-out" of tracking and opting-out of one service may mean opting-in to another. For example, Criteo is one of many services which require setting an opt-out cookie, and state in their policy that: "if your browser settings prevent the use of...third party cookies in general, the choice mechanisms offered by the platforms above will not operate properly." Conversely, Oracle instructs users to turn off third-party cookies to opt-out of tracking, which would have the effect of disabling the Criteo opt-out: "Oracle does not uniformly process do-not-track signals from browsers. However, you may prevent Oracle from collecting Interest Segments using Cookies on a browser by blocking third-party Cookies in that browser." Thus, Criteo and Oracle have policies which are fundamentally incompatible on a technical level. However, if both Criteo and Oracle interpreted "Do Not Track" as an opt-out, this contradiction could be easily resolved.

<sup>4</sup>It is important to note that the F.T.C. also stated that "work remains to be done on Do Not Track" in 2012, and such work remains unfinished in 2018.[43]

- Do third-party privacy policies mention and respect Do Not Track signals?

Questions regarding **Security** are as follows:

- What percentage of websites force encrypted connections?
- What percentage of third-party requests are encrypted?

### 3 METHODOLOGY

While the above research questions are relatively limited in scope, answering them is a non-trivial task and new methods have been developed for this study. The first task required is to determine the third-parties which collect data on a given website and if the data transfer is secure. For this task, the `webxray` software platform is used to monitor third-party network traffic generated by loading a given web page and attribute such traffic to the entities which receive the data. Second, website privacy policies must be identified, extracted, and audited. For this task, a new module for `webxray`, named `policyxray`, has been developed. Although Cranor et al have previously automated analysis of financial website policies [10], `policyxray` is the first tool capable of auditing disclosure of specific third-party data flows in website privacy policies and represents a step forward in the automation of privacy policy analysis. Finally, the relevant policies of third-party data collectors must be selected using a manual process. These steps are described below.

#### 3.1 `webxray`

`webxray` is a software platform which measures data leakage to third-parties when loading a given website. `webxray` leverages an extensive hand-curated library which attributes ownership of third-party domains to the services and corporate entities which control them. `webxray` has previously been used in academic research [17, 24, 25, 35] and the `webxray` attribution library has been used to augment findings in other platforms such as OpenWPM [12, 38].

To use `webxray`, one must first generate a list of web pages which are then loaded in a web browser. During page loading, HTTP element request and receive events are monitored. To determine privacy leakage, third-party requests are identified by comparing the domain of the page (e.g. “example.com”) to the domain of the request (e.g. “tracker.com”). Sub-domains are ignored so that a request to a domain such as “images.example.com” is not recorded as a third-party. There is no purely automated mechanism to disambiguate between site-specific sub domains and country-specific sub-domains (e.g. “example.co.uk”), so the Mozilla Public Suffix list is used for this task.<sup>5</sup>

Once third-party domains are identified, `webxray` searches for them in an internal database of domain ownership. The `webxray` database is the product of years of detective work as automated tools such as `whois` are unable to reveal the owners of anonymously registered domains. The process for determining domain ownership is often laborious, but focused human attention produces results not currently achievable by purely machine-driven approaches.<sup>6</sup> The `webxray` attribution database has been modified for this project to reveal the hierarchy of ownership which connects a service to a parent company. For example, `webxray` is able to determine that

<sup>5</sup>See <https://publicsuffix.org> for additional details.

<sup>6</sup>For example, the owner of one domain was only determined after locating obscure developer documentation.

the domain “convertro.com” is owned by Convertro, which is a subsidiary of Aol, which is a subsidiary of Oath, which in turn is owned by American telecommunications giant Verizon.

`webxray` currently supports both the Google Chrome browser and the PhantomJS headless browser. Chrome has the benefit of being the same browser many users employ and is suitable for small volumes of pages. Due to non-trivial resource requirements and instability when many instances are run in parallel, Chrome is poorly suited for large volumes of pages.<sup>7</sup> For this study, the headless browser PhantomJS has been used. On a suitably robust machine, 64 parallel instances of PhantomJS can be easily run.

A computer located at an academic institution in the United States is used to conduct measurement. Using a computer on a university IP block produces better measures than using a cloud hosting provider such as Amazon Web Services due to the fact that IP addresses from cloud hosts are often blocked as they may be used for site scraping and click-fraud. A major strength of this study is that a cloud service is not used for measurement tasks.<sup>8</sup>

#### 3.2 `policyxray`

`policyxray` is a newly developed module for `webxray` which extracts privacy policies and audits their content for disclosure of the specific third-parties which collect data on a given page. It is the first tool designed to audit observed third-party tracking in website privacy policies and represents the most significant contribution of this study.

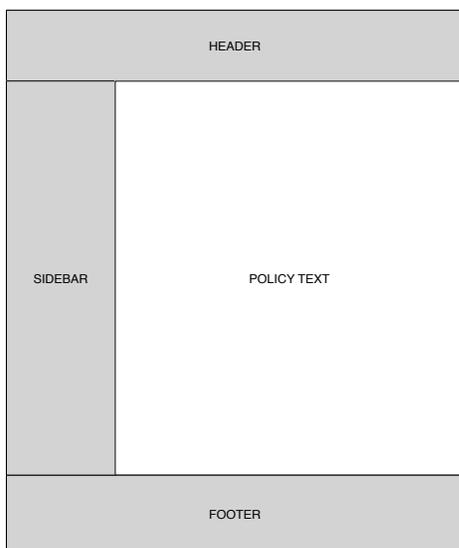
`policyxray` relies on a modification to the `webxray` software which facilitates the harvesting of privacy policy links. When `webxray` loads a page, it extracts all of the links on the page. The text of each link is evaluated to see if it contains the sub-string “privacy policy”. The first such mention is recorded as the policy link and searching stops. If there is no match, the following strings are searched for in order: “privacy”, “terms of service”, “terms of use”. Given that policy links are usually found in the footer of a page, links are evaluated in a bottom-to-top order relative to page layout.

When `policyxray` is run, it attempts to load the URL corresponding to a given site’s privacy policy. Next, `policyxray` attempts to extract the policy text from the page so that it may be evaluated independently of other page elements such as sidebars or footers. This is necessary because social media companies such as Facebook and Twitter are often mentioned in the text of a footer link, but may not be present in the policy. Figure 1 illustrates an example of how policy text differs from page text.

Extracting policy content from a web page is a difficult problem given variations in website coding styles. To overcome this, the `Readability.js` Javascript library is used. `Readability.js` is an open-source project maintained by Mozilla which provides an automated method for extracting page content by removing boilerplate sections such as page headers and footers. In order to leverage `Readability.js`, `policyxray` loads a page with either Chrome or PhantomJS, injects the `Readability` script into the page, executes it, and strips any remaining HTML elements from the text.

<sup>7</sup>This is true even when running Chrome in headless mode.

<sup>8</sup>For example, the Google Scholar webpage is easily accessible from a university IP, but loading the same page from a cloud service IP address results in a block.



**Figure 1: Footer blocks frequently include the names of social media companies. Such mentions may generate a false positive that the company is mentioned in the policy; therefore policyxray extracts only the policy text of a page.**

Readability uses a measure of “link density” to determine if a given page section is likely to be a navigational block, and therefore not a part of the article (see [19] for an extended discussion of link density). An earlier incarnation of the Readability.js library was produced by Arc90 Labs; research from 2010 found that version of Readability performed with 95.90% accuracy [37]. Readability is in active development and both Firefox and Safari use it for their “Reader View” features.

To verify that the extracted policy text is in fact a privacy policy, a basic sanity check is done to verify that the page title or text contains the strings “privacy” or “cookie”. Furthermore, manual inspection of a random sample of harvested text found that 100% of samples contained *only* policy text (95% confidence with +/- 5% interval).

The main purpose of policyxray is to determine if the parties identified collecting data on a given page by webxray are disclosed in that page’s privacy policy. To do so, the name of each domain owner is searched for in the policy text, if it is found, it is counted as disclosed. If a given owner is not found, policyxray recursively searches for mention of any parent organizations. Thus, if “convertro.com” is found on a page by webxray, policyxray will search for the strings “Convertro”, “Aol”, “Oath”, and “Verizon”. In cases where there may be variations in the name of a service, such as “DoubleClick” and “Double Click”, both are searched for. This process allows for a purposefully inclusive approach to auditing disclosure and is designed to give as many opportunities as possible for disclosure to be observed.<sup>9</sup>

<sup>9</sup>The company “Inform” was excluded from analysis due to the fact the word “inform” appears with high frequency independently of the company being disclosed. Including “Inform” vastly skews overall findings.

To determine if the “Do Not Track” (DNT) standard is mentioned and respected in a policy, the string “do not track” is searched for in the policy text. This step is easy to automate, but determining if the string is in reference the DNT standard, and if the choice signal is respected, is a difficult task. For this study, a random sample of policies with a match on “do not track” are manually evaluated to determine if the match is a reference to the DNT standard, and if so, if respect for user choice is clear.

Finally, policies are evaluated to determine how difficult they are to read and the time needed to read them. In regards to reading difficulty, the well-established Flesch Reading Ease metric is used. This metric is a means of measuring the difficulty of reading a text written in the English language. In regards to time taken to read a policy, this study adopts the approach of McDonald and Cranor who “assumed an average reading rate of 250 words per minute” [28].

### 3.3 Selection of Third-Party Privacy Policies

There are three main reasons a third-party may be collecting data on a given website. First, the party may be a Content Delivery Network (CDN). In this case, data collection may be viewed as incidental and largely outside the scope of notice and choice. Second, the party may be a service used for Distributed Denial of Service (DDoS) mitigation and data collection may also be viewed as incidental. In the final case, a third-party may be collecting user data for audience tracking, online advertising, data brokerage, or other tasks which require processing and storing data related to the behavior of specific users.<sup>10</sup>

Once webxray has produced a ranked report of the third-parties most frequently found in the studied population of sites, those parties which are CDNs and DDoS mitigation services are excluded from further consideration. For the remaining parties, the most salient privacy policy is chosen manually. The considerations in this process are first to choose the policy which is most applicable to third-party data collection, is written in English, and if there are policies for more than one country, the U.S. policy is used to reflect the location of the machine being used for the study. Once the policies are selected, it is possible to isolate the policy text, identify Do Not Track clauses if present, and evaluate readability.

### 3.4 Limitations

While the methods perform well at scale, they are not without limitation. First, because PhantomJS was used as a browser, requests for Flash elements may be missed and thus a given company collecting data will not be identified. Prior research has observed that PhantomJS may not successfully load some pages [12], but user-agent randomization greatly reduces this issue. Likewise, due to rapid ingestion of pages, it is possible that the IP address used for collection be black listed for appearing to be a “bot” and pages and elements will not load.

Second, due to the nature of real-time ad bidding, websites which rely on advertising will likely expose users to different parties on each page load depending on what parties win a given auction. Thus, for sites with advertising, loading pages a single time will produce an under-count of the number of parties which may collect

<sup>10</sup>The term “natural persons” may also apply here.

user data on the site. However, because webxray’s database of domain ownership primarily contains major ad networks rather than small clients, and policyxray only searches for identified parties, variability in the long-tail of trackers may not have an outsized effect on overall findings related to disclosure. Nonetheless, it is important to point out that the number of parties being searched for is *fewer* than the total number of parties present.

Third, although Readability.js is used by major browsers and has been tested in prior research, it is not perfect and it is possible portions of a policy may not be extracted. If such portions contain the only mention of a given third-party, that will produce a false negative. Conversely, if extraneous non-policy text is included, and that text includes the a mention of a given third-party, it may produce a false positive. However, as noted above, a sample of collected policies detected no such issues.

Finally, due to the fast pace of ownership changes in the online advertising market, it is possible that some parties may have new parents or subsidiaries which are not yet reflected in the webxray database. While the above limitations may impact findings, this study nonetheless represents the first attempt to perform the task of auditing the disclosure of third-party data flows in website privacy policies at scale.

## 4 RESEARCH FINDINGS

In October 2017, a computer based at a United States academic institution is used to scan one million popular websites as identified by the Alexa company using the webxray software platform. Of these pages, 938,093 are successfully loaded and privacy policy links for 248,029 pages are extracted. policyxray is used to extract 184,897 unique policies corresponding to 207,000 sites. The number of policies is lower than the number of sites because sites owned by a single entity often share policies. Of the most prevalent third-parties receiving user data, 25 are selected for their pertinence to the study of notice and choice and their privacy policies are extracted for analysis.

Findings shed light on the general state of tracking on popular websites, the nature of the third-parties collecting user data, rates of disclosure for third-party data flows, the complexity and length of privacy policies, respect for the “Do Not Track” (DNT) standard, and the security practices used by popular websites and third-party data collectors. Taken as a whole, findings demonstrate that there is poor disclosure of third-party data collection, policies are difficult and time consuming to read, DNT is rarely respected, and security practices are suboptimal. The follow sections address these findings in detail.

### 4.1 Current State of Web Privacy

Prior work has investigated the state of tracking on the Alexa top one million websites in both 2015 and 2016 [12, 24]. It is therefore useful to briefly provide an overview of the current state of web tracking in order to contribute to the historical record.

Of the 938,093 pages which loaded successfully, 91.27% initiate a request to download a third-party page element, thereby potentially exposing users to cross-site tracking. Pages which initiate a third-party request expose users to an average of 10.89 unique domains per page-load. The number of third-party requests for the top 10,000

**Table 1: Third-Party Prevalence, SSL Use, and First-Party Disclosure**

†Denotes Company has Consumer Services

Company	% Tracked	% SSL	% Disclosed
Google †	82.81	80.35	38.29
Facebook †	33.37	91.61	17.50
Twitter †	12.26	90.43	10.74
AppNexus	11.97	59.98	0.44
Oracle	11.21	41.51	3.72
Adobe †	10.14	70.48	5.77
Oath †	9.67	57.64	4.42
The Trade Desk	7.38	56.49	0.12
Axciom	7.10	34.21	0.26
Rubicon Project	6.68	71.62	0.12
OpenX	5.78	52.50	0.29
Lotame	5.71	29.82	0.29
IPONWEB	5.64	66.11	0.07
Casale Media	5.05	63.74	0.05
Criteo	4.93	62.26	2.75
Neustar	4.78	40.05	0.04
PubMatic	4.61	54.27	0.19
Media Math	4.60	56.23	0.04
Microsoft †	4.57	72.27	12.56
comScore	4.57	53.42	1.74
Nielsen Online	4.03	41.41	0.35
AdForm	3.96	50.71	0.88
New Relic	3.94	97.18	0.60
Quantcast	3.71	46.01	1.46
Rocketfuel	3.65	59.83	0.10

sites is 20, whereas for the bottom 10,000 it is 10 (see Figure 3 for detail). 70.60% of page loads result in the setting of a third-party cookie, and pages with third-party cookies have an average of 11.24 distinct cookies per-page. 86.84% of pages include Javascript code loaded from a third-party domain.

### 4.2 Identification of 25 Prominent Third-Party Data Collectors

As noted above, webxray uses a database of domain ownership which provides a hierarchical means to trace ownership of data collected by third-parties on the web. Table 1 shows 25 of the most prominent data collectors discovered on the Alexa top one million websites. The parties are chosen because they are primarily active in the processing and storing of the data of users rather than content hosting or DDoS mitigation. Likewise, all of the parties chosen set third-party cookies which may be used for cross-site tracking. In cases where nearly all of a single company’s data was traced back to a subsidiary, the subsidiary was selected in place of the parent. This was the case with Google (an Alphabet subsidiary) and Oath (a Verizon subsidiary).

Table 1 shows the percentage of sites which may be tracked by a given third-party. It is important to note that that for all of the companies chosen, the percentage of sites tracked is a composite measure. For example, if AOL and Yahoo are on the same site, the

site is counted once each for Aol, Yahoo, and Oath rather than twice for Oath. These composite measures provide insight into the reach of various companies. For example, Google tracks over 82% of sites, Facebook over 33%, and Twitter over 12%<sup>11</sup>. Table 1 also illustrates a long-tail distribution of third-party data collectors. The fifth place company, Oracle, tracks 11.21% of sites, a fraction of Google. Likewise, the 25th place company, Rocketfuel, tracks 3.65%, a fraction of Oracle.

Returning to the evaluation of notice, it may be assumed that if users have pre-existing consumer relationships with a company they may already be familiar with data collection practices. For example, Twitter’s privacy policy states that “We may personalize the Services for you based on your visits to third-party websites that integrate Twitter content such as embedded timelines or Tweet buttons”.<sup>12</sup> Thus, from a perspective of notice and choice, consumer services such as social media, search, and email may theoretically have provided notice of data collection independently of site policies.

Of the 25 companies examined, only six have consumer services. Therefore, for the majority of third-party data collectors there is virtually no chance users will have awareness of data collection practices due to prior interaction with a service. It is also worth noting that just because a company has a consumer service that does mean that all users who may be tracked are users of the service. For example, people who do not use Twitter, and have no reason to read Twitter’s privacy policy, may still be tracked by Twitter.

### 4.3 Disclosure of Companies in Privacy Policies

The crux of this study is an evaluation of whether or not website privacy policies provide notice of the third parties which collect data on a given site. As detailed in the methods section, `webxray` is used to determine the third-parties which collect user data on a given site. For 207,000 websites, `policyxray` is used to verify if these companies are mentioned in the site’s privacy policy. A total of 1,807,491 instances of data transmission to a known third-party are audited. It is found that only 14.80% of data transmissions to identified third-parties are disclosed. Users who read website privacy policies are therefore very unlikely to be notified of the parties which collect their data.

While the overall rate of disclosure is low, it is not uniform across parties. As Table 1 shows, transfers to Google are disclosed in 38.29% of cases. While over 60% of transfers to Google are *not* disclosed, there is clearly a strong possibility users may learn of data transfer either through a site policy, or through Google’s own policies. For companies with consumer services, disclosure is lowest for the Oath group with 4.42%. Again, because Aol and Yahoo are Oath subsidiaries it is possible users are notified via consumer policies. For all companies with consumer services, the average rate of disclosure is 14.88%.

Among the 25 prominent third-parties inspected, disclosure for non-consumer services is sharply lower. For the 19 services which most users are likely unaware of, the average rate of disclosure is

<sup>11</sup>It is interesting to note that Twitter is down from 18% in May 2014, and this is a trend worth exploring if the social network declines in relevance [24].

<sup>12</sup><https://twitter.com/en/privacy>

**Table 2: Third-Party Privacy Policy Characteristics**  
†Denotes DNT Mention  
‡Denotes DNT Partially Respected

Company	Word Count	Reading Ease
Google	2773	39.67
Facebook	2701	48.94
Twitter	3799	35.1
AppNexus	3901	43.22
Oracle †	4844	29.18
Adobe	1700	29.08
Oath †	2461	35.61
The Trade Desk †	5731	39.06
Acxiom	881	26.61
Rubicon Project	720	37.84
OpenX †‡	3345	35.31
Lotame †‡	3150	29.48
IPONWEB	947	29.48
Casale Media	1301	25.90
Criteo †	3287	38.25
Neustar	5903	31.31
PubMatic †	4360	18.42
Media Math	4794	39.16
Microsoft †	25367	40.89
comScore †	873	35.27
Nielsen Online	1566	42.41
AdForm	2134	26.85
New Relic	4150	43.06
Quantcast	2924	40.79
Rocketfuel	3445	35.10
Average	3882	35.48

less than 1%. Simply put, if a user does not have a pre-existing consumer relationship with a third-party there is virtually no chance they will learn of these parties by reading privacy policies.

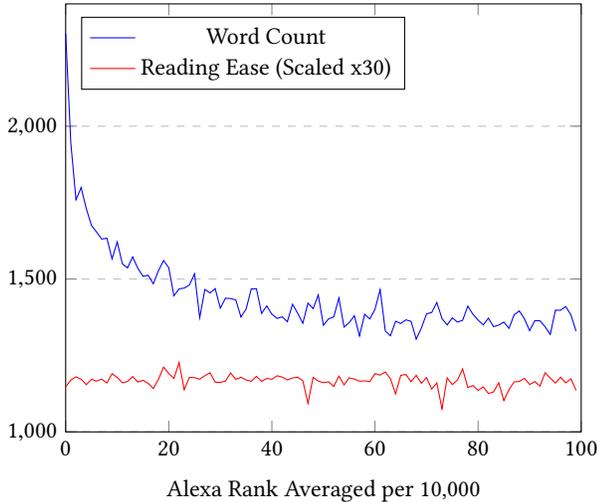
### 4.4 Readability of Policies

Beyond mere mention of the parties collecting user data, privacy policies for websites may detail a host of other issues related to data storage, retention, and use. However, in order for this information to be useful, it must be understood by most users. While the degree to which a given text is understandable relies on a host of factors ranging from a given user’s literacy to familiarity with the minutia of data protection regulations, it is possible to use the well-established Flesch Reading Ease (FRE) metric to evaluate the difficulty of reading a given text.

FRE scores range from 0-100, with lower scores indicating the text is more challenging to read. Given that it is difficult for individuals to consent to contracts which they cannot understand, the U.S. state of Florida requires that insurance policies are written in a way which generates “a minimum score of 45 on the Flesch reading ease test”.<sup>13</sup>

<sup>13</sup>Florida Statutes Section 627.4145 - Readable Language In Insurance Policies. (Fla. Stat. §627.4145)

**Figure 2: Sites ranked higher in Alexa tend to have longer policies (blue), but Flesch Reading Ease (red) shows much less variation.**



Analysis of website privacy policies reveals that they have an average FRE score of 39.83.<sup>14</sup> Thus, if an average website privacy policy were an insurance policy in the state of Florida, it would not pass basic legal requirements.

When turning attention to the privacy policies of the 25 prominent data collectors detailed in Table 2, it is found that the average readability score, 35.48, is lower than that for website policies. Again, if these policies were for insurance in Florida, rather than online privacy, they would not be valid.

There is significant variation between various third-party policies. Facebook has the most readable policy, with a score of 48.94; at the bottom end, PubMatic has a score of 18.42, which makes it more difficult to read than an article in the *Harvard Law Review* [29].

#### 4.5 Time Required to Read Policies

In 2008, McDonald and Cranor posed an intriguing question: “If website users were to read the privacy policy for each site they visit just once a year, what would their time be worth?” [28]. By analyzing the “word count of the 75 most popular websites”, they determined that the “national opportunity cost” of reading privacy policies would be \$781 billion dollars. The present study updates McDonald and Cranor’s findings, vastly expands the number of policies studied, and introduces a new composite measure to determine the time taken to read *both* first- and third-party policies.

Across policies for 207,000 sites, the average number of words per-policy is 1,404. Using “an average reading rate of 250 words per minute”, an average website policy would require 5.6 minutes to read [28]. This is lower than the average of 10 minutes found by McDonald and Cranor in 2008. This is due to the fact that policies for

<sup>14</sup>This level of difficulty shows little variation relative to the Alexa rank of a given site (see Figure 2).

sites ranked higher by Alexa tend to have longer policies, whereas the majority tend to be significantly shorter (see Figure 2).

One possible reason policies for highly-ranked sites are longer may be that they are written by teams of lawyers who create detailed, custom policies. In comparison, low-ranked sites likely do not have the resources needed to generate complex policies. Nevertheless, it appears that when looking at the larger population of sites, the time requirements of reading an average website policy could be much lower than earlier work estimated. However, this may not be the case.

As the findings presented thus far make clear, a user visiting a given website is subject to *many* policies: those of the website as well as the third-party data collectors. As Table 2 details, the average length of a third-party privacy policy is 3,882 words and would require 15.5 minutes to read, nearly three times a website privacy policy.

Because a single site may expose users to several policies in tandem, it is possible to calculate the total time needed to read *all* applicable policies of a given site. Considering that the total minutes to read applicable site policies is the sum of the word count of the first-party policy ( $WC_{fp}$ ) and the word count of all third-party policies ( $WC_{tp}$ ), divided by 250, the following formula may be derived:

$$t_{mins} = \frac{WC_{fp} + \sum WC_{tp}}{250}$$

Applying this formula reveals that the total time needed to read all applicable policies for a given site is 84.7 minutes on average. This calculation does *not* take into account that users may not need to re-read a third-party policy on every site they view, and is only applicable to the first site in which the total set of policies is encountered.

As previous sections detailed, there is a low probability users will be aware of third-party policies to start with, and it is highly unlikely any user would have the ability to locate relevant policies, let alone the time to read them. Thus, the assertion is not that that users actually spend over an hour reading policies, rather this finding underscores that the notice and choice regime is fundamentally untenable when the full range of policies is considered.

#### 4.6 Respect for User Choices

As the above findings have made clear, the likelihood of a user receiving notice of the third-parties receiving their data by reading a site’s privacy policy is remarkably low. However, it is possible that users could express their choices to control data collection in a way which would not require user notification. The “Do Not Track” (DNT) mechanism accomplishes this task and is a means for users to communicate their desire not to be tracked to parties receiving HTTP requests.

DNT is a setting available in all major web browsers and is easy for users to enable. The U.S. Federal Trade Commission has been highly supportive of the standard and encouraged its development [43]. According to the technical specification, DNT provides a “means of allowing users to express their preferences about tracking, including to opt out of tracking some or all of the time” [27].

DNT may be viewed as a polite request and there is no technical mechanism to force compliance on the part of data collectors.

Rather, data collectors must commit to respect the expressed choice signal in their policy documents. This study is the first to examine support of DNT for both websites and third-party data collectors at scale.

Across the population of website policies analyzed, 8% contain the string “do not track”. A manual analysis of a sample of policies determines if the string is in reference to DNT, and if so, if DNT is honored. It is found that 15% of instances of the “do not track” phrase are not in reference to DNT (e.g. “we do not track users” was a common phrase that is not DNT-related per se). Thus, when rounding to the nearest integer, 7% of all policies discuss DNT.

Of policies mentioning DNT, 77% explicitly do *not* honor it. For example, one representative policy stated that “we will not disable tracking technology that may be active on the Sites in response to any ‘do not track’ requests that we receive from your browser.”<sup>15</sup> A policy for a government website in Arkansas specifies that “While the United States Federal Trade Commission has endorsed DNT, our Sites do not currently support DNT codes.”<sup>16</sup>

Only 23% of policies mentioning DNT contain a clear commitment to honor a user’s DNT preference. One such commitment is found in the following statement: “We honor do not track signals and do not track, plant cookies, or use advertising when a Do Not Track (DNT) browser mechanism is in place.” However, it is important to reiterate that such commitments are voluntary and difficult to audit.

Given that a single website may not have the ability to “track” users between sites, the language of DNT may not be fully applicable. However, for the third-parties which track users across the web, DNT has particular salience. Furthermore, whereas small website operators may be ignorant of the DNT standard, major third-party data collectors are well aware of it and employ lawyers with expertise in data protection regulations.

Despite this awareness, only nine of 25 data collectors mention the DNT standard in their privacy policies. As with first-party disclosures, the majority of these mentions are to specify that DNT is ignored. Only two third-parties, OpenX and Lotame, offer qualified support for DNT. Lotame’s policy represents the best respect for the spirit of DNT: “If Lotame receives a ‘Do Not Track’ signal from any browser other than Internet Explorer, Lotame will implement an opt-out.” Twitter was previously the largest party to respect DNT, but has recently stopped doing so.

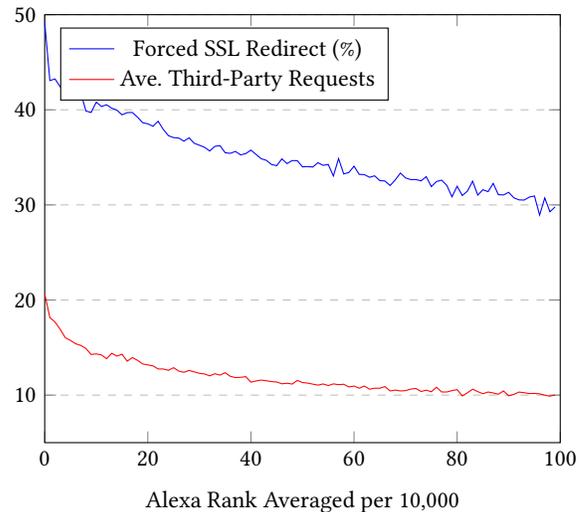
#### 4.7 Security Practices

In addition to notice and choice, guidelines for online advertising often include provisions for ensuring data security. U.S. Federal Trade Commission online advertising guidelines from 2009 assert that “Any company that collects and/or stores consumer data for behavioral advertising should provide reasonable security for that data”[42]. Likewise, the “Self-Regulatory Principles for Online Behavioral Advertising” authored by industry trade group Internet Advertising Bureau dictate that: “Entities should maintain appropriate physical, electronic, and administrative safeguards to protect the data collected and used for Online Behavioral Advertising purposes” [4].

<sup>15</sup><http://www.emaworld.com/privacy/>

<sup>16</sup><http://www.arkansas.gov/policies/privacy-policy>

**Figure 3: Higher ranked sites force SSL more often (blue), but also initiate more third-party requests (red).**



In the context of third-party data transfer on the web, there are two main technical factors involved in protecting user data: storage encryption and transport encryption. Storage encryption applies to how data is protected once it reaches its destination and protects against unauthorized parties reading data after it has been received and processed. Transport encryption refers to the process by which data is encrypted as it is transferred over the Internet and protects against network adversaries reading the data.

It is impossible to verify if third-parties collecting user data are employing sufficient transport encryption without an independent auditing body. At present, no such body provides publicly-available reports of security practices. However, it is possible to determine if transport encryption is being used by examining the network traffic generated when loading a page in order to determine if connections are made utilizing Secure Sockets Layer (SSL) connections.

While transport encryption adoption has been increasing, there is still a large volume of unencrypted traffic which places user data at risk of interception. Of all pages examined, 35.14% redirect users to an SSL-secured HTTPS page after being requested via an HTTP request. As Figure 3 illustrates, higher-ranked websites force SSL connections more often. In terms of page content, 52.25% of all element requests are encrypted. However, there is significant variability in encryption between first- and third-party requests: first-party are encrypted in 35.52% of cases compared to 66.82% for third-parties.

The above findings suggests that third-parties may have superior data security practices. However, as Table 1 illustrates, there is huge variability in the encryption practices of the 25 examined third-party data collectors. Facebook, Twitter, and New Relic all encrypt over 90% of requests. In contrast, Oracle, Acxiom, Lotame, Neustar, Nielsen Online, and Quantcast encrypt fewer than half of all requests. This wide variability underscores how self-regulation produces wildly differing practices among data collectors and suggests that clear standards should be adopted and enforced.

## 5 PRIOR WORK

This study builds directly on three established lines of research, those investigating web privacy and third-party tracking, those addressing privacy policy usability, and those addressing legal and normative objections to notice and choice.

Web privacy and third-party tracking is an active area of research. Krishnamurthy and Wills conducted several early censuses of third-party tracking [22]. More recently both Libert and Englehardt and Narayanan have investigated tracking on the Alexa one million most popular sites [12, 24]. Several studies have looked at the longitudinal evolution of third-party tracking on the web [21, 23], defenses against tracking [33], the economics of online tracking [7, 16], tracking on smartphone apps [45, 47], and the prevalence of different tracking methods [1, 6, 8]. Beyond general censuses, many studies have looked at specific implications of third-party tracking: Libert has investigated tracking on health-related webpages [25], Englehardt et al have investigated the nexus between web tracking and state-sponsored surveillance [13], and Hauschke has detailed tracking on library websites [17].

Numerous studies have investigated usability aspects of the notice and choice policy framework. McDonald and Cranor investigated the time required to read policies and the overall cost to national productivity [28]. Likewise, McDonald et al have conducted comparative work on privacy policies [29]. Another study from Cranor et al “automatically evaluated 6,191 U.S. financial institutions’ privacy notices” [10]. Reidenberg et al investigated the degree to which users have trouble understanding policies [32], while Acquisti and Grossklags have found that “even if individuals had access to complete information...they might still deviate from the rational strategy” [3]. Wilson et al have evaluated the feasibility of crowdsourcing annotation of privacy policies [46]. Komanduri et al found that members of online advertising industry groups the Network Advertising Initiative (NAI) and Digital Advertising Alliance (DAA) did not always follow their own notice and choice guidelines [20].

Legal and normative scholars have extensively addressed fundamental issues with the notice and choice policy regime. Solove has stated that the regime “cannot achieve the goals demanded of it, and it has been pushed beyond its limits” [36]. Cate has criticized the fact that the FTC has chosen not to regulate privacy practices, but instead “focused virtually all of its...efforts on getting websites to post privacy policies and its enforcement efforts on suing website operators when they fail to follow those policies” [9]. Barocas and Nissenbaum have observed that “users who are subject to [online tracking] confront not only significant hurdles but full-on barriers to achieving meaningful understanding of the practice and uses to which they are expected to be able to consent” [5]. Rotenberg has asserted that the notice and choice paradigm is “a relatively recent creation of the U.S. marketing industry” and is at odds with internationally recognized data protection frameworks [34].

## 6 CONCLUSION

This study has shown empirically for the first time that the notice and choice policy regime fails to notify users reading privacy policies of the parties which collect their data. It also demonstrates that the time burden for reading *both* first- and third-party policies

is unmanageable. Furthermore, the “Do Not Track” mechanism is rarely mentioned in privacy policies, and when it is, it is usually to specify the expressed user choice is ignored. Finally, while implementing SSL encryption is an easy means of ensuring transport security, there is uneven support across the third-parties which collect user data on the web.

Private-sector technologists routinely assert that privacy is a modern invention which may be viewed as an “anomaly” [18]. Such a viewpoint is not only ignorant of widely documented instances where privacy has been endorsed as a foundational social value by numerous ancient cultures [2, 30], it is also ignorant of the *recent* history of data protection regulations around the world [15]. Furthermore, decades of survey research have consistently demonstrated that online privacy is valued by the public [11, 14, 26, 40, 41]. Thus, the true anomaly may be that a massive sector of the global economy has been regulated by a fundamentally broken approach which is sharply at odds with public desires.

## 7 ACKNOWLEDGEMENTS

I would like to thank the reviewers for their helpful comments, Reuben Binns and Nataliia Bielova for assistance with the abstract, and my dissertation committee who oversaw earlier versions of this work (Victor Pickard, Jonathan M. Smith, Michael X. Delli Carpini, Joe Turov, Guobin Yang). Jonathan M. Smith deserves additional thanks for providing the computational resources needed for this project well beyond my dissertation defense.

## REFERENCES

- [1] 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, 674–689.
- [2] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.
- [3] Alessandro Acquisti and Jens Grossklags. 2005. Privacy and rationality in individual decision making. *IEEE Security & Privacy* 2 (2005), 24–30.
- [4] American Association of Advertising Agencies, Association of National Advertisers, Council of Better Business Bureaus, Direct Marketing Association, and the Interactive Advertising Bureau. 2010. *Self-Regulatory Principles for Online Behavioral Advertising*. (2010).
- [5] Solon Barocas and Helen Nissenbaum. 2009. On notice: The trouble with Notice and Consent. In *Proceedings of the Engaging Data Forum: The First International Forum on the Application and Management of Personal Electronic Information*.
- [6] Muhammad Ahmad Bashir, Sajjad Arshad, William K Robertson, and Christo Wilson. 2016. Tracing Information Flows Between Ad Exchanges Using Retargeted Ads.. In *USENIX Security Symposium*. 481–496.
- [7] Ceren Budak, Sharad Goel, Justin Rao, and Georgios Zervas. 2016. Understanding emerging threats to online advertising. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. ACM, 561–578.
- [8] Aaron Cahn, Scott Alfeld, Paul Barford, and S Muthukrishnan. 2016. An empirical study of web cookies. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 891–901.
- [9] Fred H Cate. 2006. *The Failure of Fair Information Practice Principles*. In *Consumer Protection in the Age of the 'Information Economy'*. Routledge.
- [10] Lorrie Faith Cranor, Pedro Giovanni Leon, and Blase Ur. 2016. A large-scale evaluation of US financial institutions’ standardized privacy notices. *ACM Transactions on the Web (TWEB)* 10, 3 (2016), 17.
- [11] Lorrie Faith Cranor, Joseph Reagle, and Mark S Ackerman. 2000. *Beyond concern: Understanding net users’ attitudes about online privacy*. Cambridge, MA: MIT Press.
- [12] 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*. ACM, 1388–1401.
- [13] 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. International World Wide Web Conferences Steering Committee, 289–299.
- [14] Oscar H Gandy. 2003. Public opinion surveys and the formation of privacy policy. *Journal of Social Issues* 59, 2 (2003), 283–299.
- [15] Robert Gellman. 2016. Fair information practices: A basic history. *Self Published* (2016).
- [16] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. 2013. Follow the money: understanding economics of online aggregation and advertising. In *Proceedings of the 2013 conference on Internet measurement conference*. ACM, 141–148.
- [17] Christian Hauschke. 2016. Third-Party-Elemente in deutschen Bibliothekswebseiten. *Informationspraxis* 2, 2 (2016).
- [18] Jacob Kastrenakes. 2013. Google’s chief internet evangelist says ‘privacy may actually be an anomaly’. (November 2013). <https://www.theverge.com/2013/11/20/5125922/vint-cerf-google-internet-evangelist-says-privacy-may-be-anomaly>
- [19] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 441–450.
- [20] Saranga Komanduri, Richard Shay, Greg Norcie, and Blase Ur. 2011. Adchoices-compliance with online behavioral advertising notice and choice requirements. *ISJLP* 7 (2011), 603.
- [21] Balachander Krishnamurthy and Craig Wills. 2009. Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, 541–550.
- [22] Balachander Krishnamurthy and Craig E Wills. 2006. Generating a privacy footprint on the internet. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*. ACM, 65–70.
- [23] 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 *USENIX Security Symposium*.
- [24] Timothy Libert. 2015. Exposing the Hidden Web: Third-Party HTTP Requests On One Million Websites. *International Journal of Communication* (2015).
- [25] Timothy Libert. 2015. Privacy Implications of Health Information Seeking on the Web. *Commun. ACM* (2015).
- [26] Mary Madden. 2014. Public Perceptions of Privacy and Security in the Post-Snowden Era. *Pew Research Center* (2014).
- [27] Jonathan R Mayer and Arvind Narayanan. 2011. Do Not Track: A Universal Third-Party Web Tracking Opt Out. *Internet Engineering Task Force* <http://www.ietf.org/archive/id/draft-mayer-do-not-track-00.txt> (2011).
- [28] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The Cost of reading privacy policies. *I/S: A Journal Of Law And Policy For The Information Society* 4 (2008), 543.
- [29] Aleecia M McDonald, Robert W Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. 2009. A comparative study of online privacy policies and formats. In *Privacy Enhancing Technologies*. Springer, 37–55.
- [30] National Institutes of Health, History of Medicine Division. 2002. Greek Medicine. [http://www.nlm.nih.gov/hmd/greek/greek\\_oath.html](http://www.nlm.nih.gov/hmd/greek/greek_oath.html) (2002).
- [31] Network Advertising Initiative. 2011. NAI Code of Conduct. *Network Advertising Initiative* (2011).
- [32] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia M McDonald, Thomas B Norton, Rohan Ramanath, et al. 2014. Disagreeable privacy policies: Mismatches between meaning and users’ understanding. *Berkeley Technology Law Journal* (2014).
- [33] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 12–12.
- [34] Marc Rotenberg. 2001. Fair information practices and the architecture of privacy (What Larry doesn’t get). *Stanford Technology Law Review* (2001), 1.
- [35] Yaokai Feng Danilo Vasconcellos Vargas Kouichi Sakurai Shiqian Yu, Tomohisa Ishikawa. 2017. Privacy Leakage of Job-related Information Seeking in Online Social Networks. *Information Processing Society of Japan SIG Technical Report* (2017).
- [36] Daniel J Solove. 2012. Introduction: Privacy self-management and the consent dilemma. *Harvard Law Review* 126 (2012), 1880.
- [37] Alex Spengler and Patrick Gallinari. 2010. Document structure meets page layout: loopy random fields for web news content extraction. In *Proceedings of the 10th ACM symposium on Document engineering*. ACM, 151–160.
- [38] Jeffrey Han Steven Englehardt and Arvind Narayanan. 2018. I never signed up for this! Privacy implications of email tracking. *Proceedings on Privacy Enhancing Technologies* (2018).
- [39] Martino Trevisan, Stefano Traverso, Hassan Metwally, and Marco Mellia. 2017. Uncovering the Flop of the EU Cookie Law. *arXiv preprint arXiv:1705.08884* (2017).
- [40] Joseph Turow and Michael Hennessy. 2007. Internet privacy and institutional trust insights from a national survey. *New Media & Society* 9, 2 (2007), 300–318.
- [41] Joseph Turow, Michael Hennessy, and Nora A Draper. 2015. The Tradeoff Fallacy, How Marketers are Misrepresenting American Consumers and Opening Them Up to Exploitation. *The Annenberg School for Communication, University of Pennsylvania* (2015).
- [42] United States Federal Trade Commission. 2009. FTC staff report: Self-regulatory principles for online behavioral advertising. (2009).
- [43] United States Federal Trade Commission. 2012. Prepared Statement of the Federal Trade Commission On The Need For Privacy Protections: Perspectives From the Administration and the Federal Trade Commission. (May 2012).
- [44] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*. ACM, 4.
- [45] Max Van Kleek, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J Weitzner, and Nigel Shadbolt. 2017. Better the devil you know: Exposing the data sharing practices of smartphone apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5208–5220.
- [46] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A Smith, and Frederick Liu. 2016. Crowdsourcing Annotations for Websites’ Privacy Policies: Can It Really Work?. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 133–143.
- [47] Jinyan Zang, Krysta Dummit, James Graves, Paul Lisker, and Latanya Sweeney. 2015. Who knows what about me? A survey of behind the scenes personal data sharing to third parties by mobile apps. *Technology Science* 30 (2015).