



# SSOLogin: A framework for automated web privacy measurement with SSO logins

Tien-Huy Pham  
huypt.15@grad.uit.edu.vn  
VNUHCM-UIT  
Ho Chi Minh, Vietnam

Quoc-Huy Vo  
huyvq@nii.ac.jp  
Sokendai  
Tokyo, Japan

Ha Dao  
hadao@mpi-inf.mpg.de  
MPI-INF  
Saarbrücken, Germany

Kensuke Fukuda  
kensuke@nii.ac.jp  
NII/Sokendai  
Tokyo, Japan

## ABSTRACT

Single Sign-On (SSO) enables users to access multiple websites and applications using a single set of login credentials. Undoubtedly, SSO makes it easy for users to log in to multiple websites without remembering credentials. However, it is also important to consider the potential risk of users being unaware of how the identity of their account will be utilized, the development of online tracking techniques, and any potential exchange of information with third parties without the user's knowledge. In this paper, we propose *SSOLogin*, which enables us to perform large-scale automation of website login through an SSO account. We confirm that *SSOLogin* automatically logs in to 91.8% of SSO account available websites in Tranco Top 500 sites. Next, by crawling Tranco Top 10K websites with *SSOLogin*, we show that 1,420 sites (14.2%) contain SSO login (Google and Facebook) as of July 2023, primarily on Information Technology and News/Media websites in the United States and the United Kingdom. We then shed light on the characteristics of privacy leakage of websites with SSO logins by setting up measurements in Japan. We find that 99% of websites have third-party online tracking activities, which may pose risks to user privacy. After SSO login, the target website shows an increase in third-party tracking domains. Logging in with Google adds 81 new tracking domains, while Facebook adds 33 new domains. Despite the convenience of logging in with an SSO account, it is important to be aware of the potential privacy risks associated with this practice.

## CCS CONCEPTS

• Security and privacy → Privacy protections.

## KEYWORDS

Web tracking, measurement, Single Sign-On, SSO login

## ACM Reference Format:

Tien-Huy Pham, Quoc-Huy Vo, Ha Dao, and Kensuke Fukuda. 2023. *SSOLogin: A framework for automated web privacy measurement with SSO logins*. In *Asian Internet Engineering Conference (AINTEC '23), December 12–14, 2023, Hanoi, Vietnam*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3630590.3630599>

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 the author(s) 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](mailto:permissions@acm.org).

*AINTEC '23, December 12–14, 2023, Hanoi, Vietnam*

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-0939-5/23/12...\$15.00  
<https://doi.org/10.1145/3630590.3630599>

## 1 INTRODUCTION

The methods used to track online user activities are always advancing [13]. In recent times, many websites prompt users to use their social networking identities to log in. This practice has raised concerns about online privacy and tracking. It is important that third parties may be able to obtain a user's Personal Identifiable Information (PII) without their knowledge or consent [15][16].

The use of the Single Sign-On (SSO) has become increasingly popular for authenticating website users [18]. Though it is a quick and easy way to authenticate, concerns have been raised about the privacy of Internet users [11]. Due to the complexity of the website login process, diversity in website design, and variety of web-building platforms, there is no efficient way to automatically sign in to such sites with SSO, to fill out registration forms, and to fully automate measurements at a large scale. The majority of related work measuring SSO privacy requires manual detection, execution, and registration for SSO sign-in [21, 23, 31]. Also, there are many aspects such as tracking entities that have not been addressed. Since web privacy measurement and analysis plays a core role in enhancing Internet users' understanding and knowledge [17], a general, modular, automated, and scalable framework is required to support any privacy measurement in this context. For example, a user visits a website *example.com* which provides Google login, and the user signs in to this website using their SSO account. During this process, many entities, different from the visited website *example.com*, are able to collect user information and track user activities (see also Figure 1).

In this paper, we propose a framework, *SSOLogin*, which supports the automation of website logins using an SSO account (§ 3.1). As a validation of the framework with the Tranco Top 500 websites, we show that *SSOLogin* can automatically sign in and record website contents successfully on 91.8% of the tested websites (§ 3.2). We then characterize SSO logins in the Tranco Top 10K sites (§ 4). We detected 1,420 sites (14.2%) containing SSO sign-in in the Tranco Top 10K sites as of July 2023 with our framework. Those websites are mainly Information Technology and News/Media websites in the United States and the United Kingdom. Finally, we set up a measurement to conduct online tracking analyses on the SSO login traces (§ 5.1). We find that most SSO login processes come with online tracking, which may pose risks to user's online privacy. By investigating the authentication flows of SSO logins for Tranco Top 10K sites, we point out that more than 99% of the first parties (the visited websites) send tracking requests to third parties for Google and Facebook SSO logins. These first parties' requests are sent to 959 and 943 third-party web tracking for Google SSO login and Facebook SSO login, respectively (§ 5.2). Next, we explore how websites integrate third-party request tracking into them by traceback

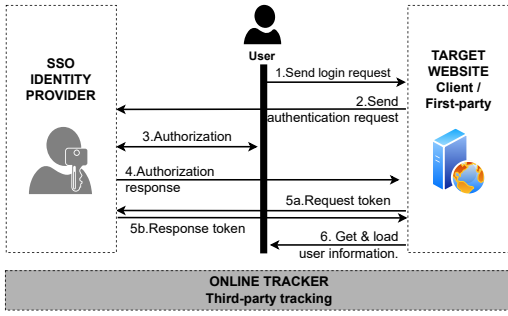


Figure 1: Basic SSO authentication workflow

method (§ 5.3). Finally, we demonstrate an increase in third-party tracking on a website after logging in with SSO (§ 5.4). Thus, the past works that did not consider SSO login largely underestimated the degree of privacy leakage.

The main contributions of the paper are as follows:

- (1) We design and implement a *SSOLogin* framework to automate SSO logins. It successfully logged in to 91.8% of SSO-supported websites in Top-Tranco 500 websites.
- (2) By crawling Tranco Top sites (up to 10,000 sites) with *SSOLogin*, we characterize privacy leakage in SSO available websites in terms of third-party tracking.

## 2 BACKGROUND AND RELATED WORKS

### 2.1 Background

**2.1.1 Web tracking.** The web relies heavily on a significant funding source, behavioral advertising. This type of advertising is tailored to individuals based on their personal information, interests, and past behaviors, also known as Online Behavioral Advertising (OBA) [1, 10, 14, 28]. By carefully monitoring an individual’s web usage across various platforms, tracking providers can gather essential information that helps advertisers better target their advertisements and attribute product sales. This process involves an extensive collection of data, including purchase history and browsing activity, all in the pursuit of achieving desired goals.

Since the Internet has revolutionized the way we interact with the world, concerns over user privacy have become a topic of many studies [12, 20, 24]. Furthermore, advertisers have increasingly adopted the practice of collecting data on Internet users. Almost every website that a user visits records, shares, and compiles data about their visit with third-party entities. The web’s omnipresence in modern life is due to its versatile design, which allows any website to pull in content from numerous sources. As a result, web developers can utilize this to build and monetize web applications that offer an immersive user experience. However, it has greatly reduced the cost of individual data collection while significantly complicating efforts to protect user privacy [17]. Web tracking and collecting personal data online can be invasive and unsettling [22, 29].

**2.1.2 Single Sign-On.** SSO is a process that allows users to log in to multiple applications and services using a single set of login credentials from an Identity Provider (IdP). Figure 1 describes the

workflow of the SSO authentication process. (1) The SSO process starts with the user sending a login request with an SSO account to the target website. (2) The target website then forwards the request to the SSO IdP, along with additional data. (3) The IdP sends a consent screen to the user, who can choose to accept or deny authorization. (4) The user’s response is then forwarded to the target website. (5) If accepted, the target website requests an access token from the IdP, which is provided along with an expiration time. (6) With the access token, the target website can access the user’s information. During this process, there are potential privacy leaks, such as the deployment of the tracker without the user’s consent (see § 5).

### 2.2 SSO logins measurements

To get a better understanding of SSO logins, some studies focus on a particular way to measure user privacy of SSO logins. We summarize the scope of their works in Table 1.

Zhou et al. [32] developed *SSOScan*, which automatically checks for single sign-on vulnerabilities in web applications. Their study detected then automatic login with a Facebook account, then diagnostic analysis of vulnerabilities found on websites. Li et al. [21] developed *OAuthGuard* which works with relying parties and uses a specific *referer* leakage detection module to detect unintentional token exchange leaks. *OAuthGuard* is a browser extension, that scans every HTTP request, filters out OAuth or OpenID Connect requests/responses, then analyzes the vulnerabilities and takes necessary protection action. Morkonda et al. [23] developed *OAuthScope*, a web tool that allows automation scan and implementation of login to the website with an SSO account, then collected data on the use of OAuth-based logins. However, *OAuthScope* only focuses on scopes and parameters of OAuth authentication, in other words, it stops at the endpoint of SSO IdP. Recently, Westers et al. [31] presented *SSO-MONITOR* for SSO detection techniques, executed the SSO and detected token exchange leaks in this process. However, fully completing the registration form, which is a crucial aspect of many privacy measurements, is currently beyond their scope.

Our work not only presents a framework that achieves high accuracy in SSO detection, execution, and registration form completion but also extensively compares a wide range of privacy measures, including tracking entity comparison leakage. Furthermore, with our approach, we enhance running speed, and increase scan and detect performance per website compared to previous studies. In addition, we also address the limitations related to different languages while keeping the framework lightweight and adaptable.

## 3 SSOLogin FRAMEWORK

In this section, we propose the *SSOLogin* framework which is capable of automating website login using an SSO account. The framework is designed to be versatile, modular, and scalable, enabling it to support practically any privacy measurement.

### 3.1 The design and implementation of *SSOLogin*

*SSOLogin* framework is composed of three main modules: (1) control manager, (2) SSO login process, and (3) data aggregator, as shown in Figure 2. We explain these modules in the subsections.

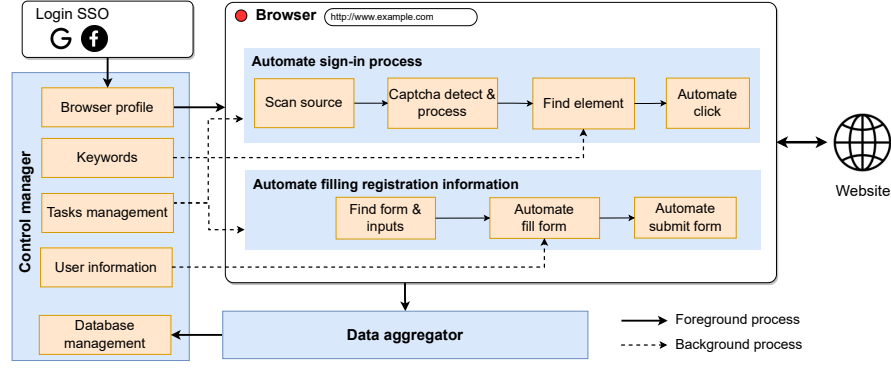


Figure 2: High-level overview of *SSOLogin*. The control manager monitors browser managers, which convert high-level commands into automated browser actions. The data aggregator receives and pre-processes data from instrumentation.

Table 1: Summary of the related works.

| Ref.                | Venue      | Year | SSO IdP    | Target           | Scope of work |         |           |             |          |         | Measure |           |          |
|---------------------|------------|------|------------|------------------|---------------|---------|-----------|-------------|----------|---------|---------|-----------|----------|
|                     |            |      |            |                  | Detect        | Execute | Reg. fill | Delineation | Security | Privacy | Request | JS script | Endpoint |
| Zhou et al.[32]     | USENIX Sec | 2014 | f          | Quantcast US 20K | ✓             | ✓       | ✓         | ⊗           | ✓        | ⊗       | ✓       | ⊗         | ⊗        |
| Li et al.[21]       | SSR        | 2019 | G          | Majestic 1K      | ⊗             | ⊗       | ⊗         | ⊗           | ✓        | ⊗       | ✓       | ⊗         | ✓        |
| Morkonda et al.[23] | WPES       | 2021 | G+ f+ A+ M | Alexa 500*5      | ⊗             | ⊗       | ⊗         | ⊗           | ⊗        | ✓       | ✓       | ⊗         | ✓        |
| Westers et al.[31]  | arXiv      | 2023 | G+ f+ A    | Tranco 10K       | ✓             | ✓       | ⊗         | ⊗           | ✓        | ✓       | ✓       | ⊗         | ✓        |
| Our work            |            |      | G+ f       | Tranco 10K       | ✓             | ✓       | ✓         | ✓           | ⊗        | ✓       | ✓       | ✓         | ⊗        |

**3.1.1 Building keyword lists.** To ensure effective scanning of web elements, *SSOLogin* primarily relies on sets of keywords. Therefore, building comprehensive sets of keywords is essential to maintain generalization amidst the growing diversity of websites. To this end, we manually visit the Tranco Top 500 websites. We discover that the user interacts with the SSO by clicking the *Login* button and *Sign in with [SSO] account* button, in which *[SSO]* is the name of SSO IdP. During this manual process, we create two distinct keyword sets tailored for these buttons: the *Login* button keywords encompass (login, sign in, account, register, and sign up), while the *[SSO]* button is primed with (sign in with [sso] and continue with [sso], [sso]).

**3.1.2 Control manager.** The control manager provides a scriptable command-line interface for controlling the distribution of commands to browser managers. It provides initialization parameters for the framework as a high-level language, such as a target site, an SSO account, browser configuration, keywords, and a *persona* containing personal information to fill in the sign-up and sign-in form. The control manager creates, monitors, passes, and converts the high-level parameters into the automated process as workflow:

- (1) Based on the configuration parameters, the control manager creates a clean browser using *Undetected-Chromedriver Selenium*<sup>1</sup>, is a Selenium WebDriver that has been optimized to avoid triggering anti-bots on website. On this clean browser, it automates login to an SSO account and saves the browser profile into a directory. This directory ensures that all browser configuration, history, and cookies are preserved for the next time.

- (2) For every input site, the control manager retrieves a previously saved browser profile and uses it to visit the site.
- (3) The control manager passes initialization parameters and runs the automated process.
- (4) Once the process is completed, the control manager waits for 90 sec to intercept data in each page from the data aggregator and saves the data to the database.
- (5) Finally, the control manager closes all browsers and releases the resources to end the process.

**3.1.3 SSO login process.** After loading the browser profile containing an SSO account from the profile directory, *SSOLogin* automates visit, detection, and log in to a website using that SSO account as workflow:

**Accept cookies banner.** A cookie consent banner is a notification that asks for user consent to allow the website to store cookies on their browser. If the cookie consent banner is not handled properly, it can prevent us from proceeding with our activities and gathering data from the website. Whereas accepting the privacy policies, web tracking is far more pervasive [19]. At this stage, we use a set of keywords from [27] as input in *SSOLogin* to automatically accept the cookie consent banners.

**Captcha verification.** We observe that many websites use Captcha for human verification, then we use the Audio Captcha function provided by Google Captcha. We record the emitted sound and use the Python library *SpeechRecognition*<sup>2</sup> to convert that sound into plain text.

<sup>1</sup><https://github.com/ultrafunkamsterdam/undetected-chromedriver>

<sup>2</sup><https://pypi.org/project/SpeechRecognition/>

**Automated SSO login process.** *SSOLogin* searches for any potential SSO login information through the Document Object Model on the page by top-down and bottom-up detection. We find out all web elements whose value in the  $\{text, aria-label, href, src, id, class\}$  attributes contains the keywords related to SSO login. We optimize the SSO login process by efficiently matching various attributes to identify all relevant elements. After finding all available elements, we automatically exclude unnecessary elements (e.g., script, hidden or wrong elements) though these text elements appear in our keywords. We only pick up the web element that is clickable and visible on the website. We also scan *iframes* on the website to make sure we do not miss any web elements belonging to SSO logins. From there, *SSOLogin* sends a command to automate clicking on this web element.

Through testing and manual inspection on many different websites, we found that the positions of the “Sign-in with [SSO]” button on websites usually have similar locations. In the most common case, after loading the site, click “Login”, then click “Sign-in with [SSO]”. Another case is that a site has a “Sign-in with [SSO]” button directly on the homepage. Furthermore, in some special sites that have a user interface design like a drop-down list of menus, the position of that button is “Login/Register” → “Login” → “Sign-in with [SSO]” or “Login” → “Sign-in with [SSO]” → Google/Facebook. Based on this observation, we find elements and automate clicking on the “Login” button and “Sign-in with [SSO]” button by using a set of pre-defined keywords (see § 3.1.1). The login process is considered successful if the website redirects to the visited website with an SSO login.

**Complete registration form.** Some websites require additional personal information from users after login with an SSO account for the first time. To handle this issue, we prepare a *persona* containing personal information identity: full name, gender, date of birth, email, phone number, address, username, password, and a random text. During the information-filling stage, our framework scans the webpage for relevant forms and identifies any empty input fields. Using the keywords and attributes associated with each field, the framework automatically fills in the required information from the persona. This process is repeated up to five times to ensure that all forms are correctly filled out, including those with multiple steps or complex input requirements.

**3.1.4 Data aggregator.** After using the control manager for the login, we use the Chrome DevTools Protocol [26] to log all web traffic. Web traffic always follows the principle of sending requests to the server and responses from the server to the client. This process mainly carries web exchange data, so the observed data will focus on these two events *request* and *response*. During the automated process, taking the network activity data, the following two network-related events are extracted:

- (1) *Network.requestWillBeSent*: triggered when a page is about to send HTTP requests.
- (2) *Network.responseReceived*: occurs when HTTP responses are available.

Besides collecting network information, we also store cookies (from both the first-party and third-party domains) and browser






local storage. In the last step, the *Control manager* saves all data in a local SQLite database and terminates the browser instance.

## 3.2 Evaluation

To evaluate whether login automation in *SSOLogin* is accurate and robust, we first test our framework to automate login using an SSO account on the Tranco top 500 websites (generated on February 13, 2023). Out of 500 sites, through manual visits to each website, we confirm that 99 websites (19.8%) were not reachable, leaving a survey list containing 402 websites. In 325 out of 402 sites with login function (including login by username or phone number or SSO), we found top five most popular SSO IdPs as 155 sites (47.7%) allow login with a Google account, 92 sites (28.3%) with a Facebook account, 91 sites (28.0%) with an Apple account, 27 sites (8.3%) sites with a Microsoft account, 24 sites with a Twitter account (7.4%).

We run *SSOLogin* for automated login with the five most popular SSO accounts: Google, Facebook, Apple, Microsoft, and Twitter<sup>3</sup>. Table 2 shows the performance of *SSOLogin* for automatic login with each type of SSO account on the Tranco top 500 websites. We define *Available* as sites that have login function by SSO account through manual inspection on each site; *Success* as sites that *SSOLogin* successfully automate login by SSO account, and *Failed* as sites that *SSOLogin* can not automate login. After comparing with manual results, *SSOLogin* is able to successfully detect and log in to these websites, with an average success rate of 91.8%. In detail, log in using Google, Facebook, Apple, Microsoft, and Twitter with success rates of 91.6%, 92.4%, 92.3%, 92.6%, and 87.5%, respectively. In addition, there are only seven unique sites on which *SSOLogin* failed to log in using an SSO account automatically. By manual verification, we find that the automated process can not pass due to technical constraints, such as required user interaction (e.g., click and hold, random questions).

**Table 2: Performance of *SSOLogin* on the top five most popular SSO IdPs of the top 500 Tranco websites.**

| SSO IdP   | Available    | Success     | Failure   |
|---|--------------|-------------|-----------|
|  | 155 / 100.0% | 142 / 91.6% | 13 / 8.4% |
|  | 92 / 100.0%  | 85 / 92.4%  | 7 / 7.6%  |
|  | 91 / 100.0%  | 84 / 92.3%  | 7 / 7.7%  |
|  | 27 / 100.0%  | 25 / 92.6%  | 2 / 7.4%  |
|  | 24 / 100.0%  | 21 / 87.5%  | 3 / 12.5% |

Compared with other approaches, we confirm that the performance of *SSOLogin* is slightly better than *SSO-MONITOR* [31] which is recently published in the community (80.2%; sample test: Tranco Top 1000 sites); Note that *SSO-MONITOR* has not supported registration form completion. In addition, our performance is close to *SSOScan* [32] in the detection process (94.2%; sample test: Quantcast Top 100 sites), but ours is much better than that in the registration form completion process (81.0%; sample test: the 973 websites that detected Facebook SSO login). From these results, we conclude that

<sup>3</sup>Note that, by analyzing the complexity of each IdPs, the design of *SSOLogin* can easily extend to other SSO logins such as Github, LinkedIn, QQ, WeChat, Weibo, and Baidu.

the performance of our framework is enough for further SSO login investigation.

**Ethical considerations:** In terms of ethics, we used a simulated SSO account to complete the authentication flows on the target websites. We are confident that our procedure did not affect their website operations. Furthermore, automating the process would significantly enhance transparency on the Internet.

## 4 WEBSITES CONTAINING SINGLE SIGN-ON CHARACTERIZATION





We first select websites that would be most appropriate for our work. We use the Tranco popular site list, which improves upon the shortcomings of the existing popular site lists: being unstable, having unreachable domain presence, and containing domains that are easily altered by an adversary [25].

To characterize websites containing SSO, we use *SSOLogin* to conduct an automatic SSO detection on the Tranco top 10K websites. We performed a crawl with the default settings of Chrome 114 browser to collect data on July 2023, with two IP addresses in Japan. Since Google, Facebook, and Apple are the most popular SSO IdPs (see § 3.2), in which Apple requires two-factor authentication, we analyze Google and Facebook IdPs in the Tranco Top 10K sites.

### 4.1 SSO usage

We first investigate the type of SSO logins on the Tranco top 10K websites to see the SSO landscape overview. We summarize our results in Table 3.

Table 3: SSO logins usage in Tranco 10k.

| SSO login   | Number of websites | Percentage |
|---|--------------------|------------|
|  (exclusively)   | 736                | 51.8%      |
|  (exclusively)   | 201                | 14.2%      |
|  &  (overlap) | 483                | 34.0%      |
| <b>Total</b>  | 1,420              | 100.0%     |

Out of 10K websites, 2,388 sites (23.88%) were not reachable during our analysis — most likely due to domain name issues, server errors, or downtime. We excluded these websites from our analysis, leaving a final dataset containing 7,612 sites. Out of these 7,612 sites, we detected 1,420 websites (18.65%) that support login with at least one IdP SSO. In total, we found 937 websites (66.0%) offering SSO support with one of the two IdPs. Within this group, Google is supported on 736 websites (51.8%) and Facebook with 201 sites (14.2%). Interestingly, 483 out of 1420 sites (34.0%) offer support for multiple IdPs (Google and Facebook).

### 4.2 How do popular sites deploy SSO?

Next, we focus on how popular websites support SSO login, because more popular sites have more impact on the privacy/security issues of many users. Figure 3a presents the Empirical Cumulative Distribution Function (ECDF) of the Tranco ranking of websites containing SSO logins. These websites are spread across the Tranco ranking. Indeed, we confirm that the plot is above the diagonal

(reference) line (for the rank less than 3000), meaning that the SSO available sites appear more in higher ranks.

### 4.3 Category of websites containing SSO

Here, we discuss the website category of websites containing SSO login. We used the FortiGuard Web Filtering [8] dataset in July 2023 for the website category classification.

Figure 3b demonstrates that SSO login is used mainly in information technology and news websites (25.28% and 14.58% in total). The percentages for business, education, and shopping are also popular (8.03%, 5.56%, and 5.21%, respectively). Moreover, we observe that there are notable similarities in the business customers of Google and Facebook SSO IdPs. Specifically, we discover that Google and Facebook have a common focus on Information Technology sites, with Google holding a larger share (31.9%) compared to Facebook (16.4%). Additionally, the two IdPs also have a significant presence in the News and Media sites market, with 14.9% for Facebook and 9.9% for Google. These findings remain consistent even for sites that utilize both Google and Facebook in the Information Technology and News and Media markets (18.8% and 21.5%, respectively).

### 4.4 Country of websites containing SSO

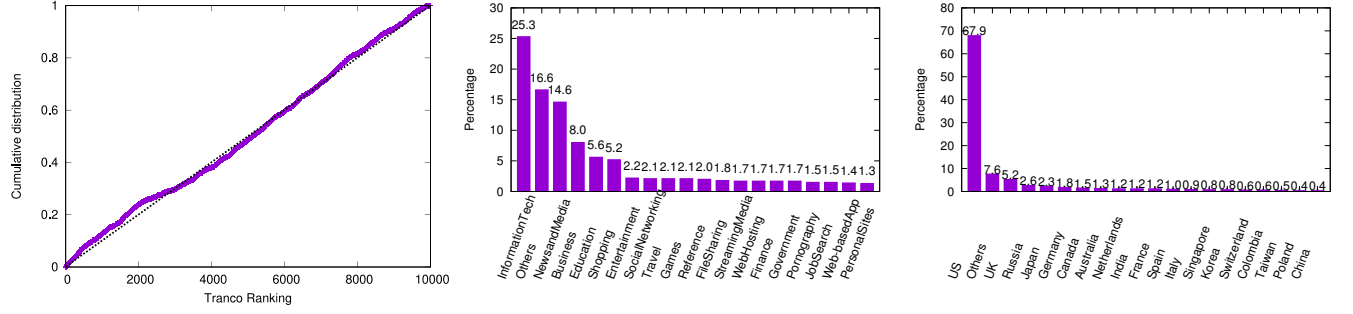
We then analyze the website country of the 1,420 websites containing SSO based on the top-level domain (TLD) and IP Geolocation. First of all, if the TLD of a website corresponds to a country (i.e., ccTLD), we attribute that website to the country. By doing that, we identify the country of 323 websites. Then, for international TLDs, we use IP Geolocation to determine the country of the 1,097 remaining websites. We are aware that, if a website uses cloud-based security, proxy, or DNS-based service, then the geolocation of the returned IP address could be unreliable for our purpose. However, this error was negligible, especially, since there are a small number of such websites containing SSO.

In addition, IP Geolocation sometimes returns incorrect results [30]. To overcome this limitation, we make a majority voting via *ip-api.com* [5], *freegeoip.app* [4], and *MaxMind* [9] to give more robustness to the Geolocation assignment. In the 16 cases of three databases returning different results or returning null, we set these websites to an unknown country. The results are shown in Figure 3c. We observe that 67.9% of websites are located in the United States, 5.2% are located in the United Kingdom, 2.6% are located in Russia, 2.3% are located in Japan, and other countries have significantly lower percentages. Furthermore, when we consider the SSO logins separately, the United States and the United Kingdom still use Facebook SSO and Google SSO more than other website countries. It is probably natural if we consider the possibility that some other countries (e.g., China) may have their own independent SSO ecosystems.

## 5 SINGLE SIGN-ON ONLINE TRACKING MEASUREMENT

Now we characterize online tracking in SSO available sites with *SSOLogin*. We complement the existing research by analyzing the effect of SSO logins on tracking requests.





(a) ECDF of the Tranco ranking of websites containing SSO logins (Tranco Top 10K sites in 2023). (b) Breakdown of websites containing SSO logins by website category (Tranco Top 10K sites in 2023). (c) Breakdown of websites containing SSO logins by website country (Tranco Top 10K sites in 2023).

## 5.1 Data collection

As described in Table 3, out of 10K websites, Google is supported on 736 websites, Facebook with 201 sites, and 483 websites offer support for both Google and Facebook IdPs. To conduct a general comparison of SSO login, we use *SSOLogin* to automate visits, login into sites using Google or Facebook accounts, then crawl each website from 483 sites supporting both IdPs. In a 10-day measurement campaign, *SSOLogin* successfully automated login 425 sites out of 483 sites (87.99%).

Then we inspect individual URLs using well-known advertising and tracking blocklists. The existing blocklists target advertisement and tracking resources and are utilized by browser extensions that intend to protect user privacy online. We consider Easylist [6] and EasyPrivacy list [7]. EasyList is a blocklist that removes ads from web pages and it is used by popular browser plugins. EasyPrivacy is an optional supplementary blocklist that removes trackers from websites. We use *Adblockparser* [3] to evaluate and identify tracking resources by matching URLs against them. To determine if a request would have been blocked by an extension utilizing these lists, we directly match the block list rules quoted above with 230,227 and 202,381 third-party requests and all requests in their request initiator chains for Google login and Facebook login, respectively. Note that *Adblockparser* has certain limitations [2], which implies that we may establish a lower bound on the extent of online tracking in this situation.

## 5.2 Online tracking in SSO schema

Firstly, we observe that most SSO login processes come with third-party tracking (see Table 4). Particularly, we detect that 100% of first parties send third-party web tracking requests for Google SSO login and 423 out of 425 first parties (99.53%) send third-party web tracking requests for Facebook SSO login. These first parties' requests are sent to 959 (by 123,831 requests) and 943 third-party web tracking (by 105,490 requests) for Google SSO login and Facebook SSO login, respectively. Overall, using Google to log in reveals user information to a slightly greater number of third-party web tracking domains compared to Facebook (Google has an average of 2.26 third-party web tracking domains per site, while Facebook has 2.21 domains per site). More interestingly, the number of third-party web tracking requests per site is large, while the number of third-party domains per site is small (Mean *third-party tracking*

*requests/third-party tracking domain/site*: 291.37/2.26 for Google and 248.21/2.21 for Facebook). This finding implies that each tracking domain will receive a lot of tracking requests on each website in SSO authentication progress, possibly from many different senders, thereby increasing the amount of online information sent with those tracking requests, which may pose risks to user privacy.

Also, we provide a breakdown of the third-party domains that received tracking from the 425 first-party senders for Google SSO login and from the 423 first-party senders for Facebook SSO login in Table 5. Interestingly, we find that when users log in to a website with a Google SSO login, 253 first parties send 2,903 tracking requests to *facebook.com* (even without the Facebook SSO login). Similarly, when users log in to a website with a Facebook SSO login, 347 first parties send 4,179 tracking requests to *google.com*. From Table 5, we can see that out of the top 10 third-party web tracking, some common third-party web tracking with user data collection functions appear such as *facebook.com* (Meta Pixel service of Facebook), *doubleclick.net*, *google-analytics.com*. This finding implies that there is a common set of third-party domains that have tracking users regardless of whether users logged in through Google or Facebook. It suggests that these third parties have established mechanisms to collect and potentially utilize user data from multiple sources, indicating a potential privacy concern. In addition, we observe that 915 third-party tracking domains are available across both SSO IdPs. Whereas there are 44 third-party tracking domains that are only available when logging in with the Google IdP process and 28 third-party tracking domains that are only available when logging in with the Facebook IdP process (see Table 6). Our result shows that Google and Facebook are the favorite online user-tracking services. In addition, many other tracking domains also receive a considerable number of tracking requests from first-party

Table 4: Breakdown of online web tracking on SSO logins.

| Metric                                | G       | F       |
|---------------------------------------|---------|---------|
| Number of sites                       | 425     | 425     |
| Number of 3rd-party tracking requests | 123,831 | 105,490 |
| Unique 1st party→3rd party tracking   | 425→959 | 423→943 |
| Mean 3rd-party tracking requests/site | 291.37  | 248.21  |
| Mean 3rd-party tracking domain/site   | 2.26    | 2.21    |

websites. This shows that there are now more tracking service providers available beyond the well-known providers.

**Table 5: Breakdown of top 10 third-party web tracking domains on SSO logins. Percentages are given out of a total of unique first parties domains for each SSO IdP.**

| #  | Third-party           | First-parties |              |
|----|-----------------------|---------------|--------------|
|    |                       | G(# / %)      | F(# / %)     |
| 1  | google.com            | 425 / 100.00% | 347 / 82.03% |
| 2  | doubleclick.net       | 383 / 90.12%  | 381 / 90.07% |
| 3  | googletagmanager.com  | 373 / 87.76%  | 373 / 88.18% |
| 4  | google-analytics.com  | 361 / 84.94%  | 358 / 84.63% |
| 5  | google.co.jp          | 335 / 78.82%  | 335 / 79.20% |
| 6  | facebook.com          | 253 / 59.53%  | 389 / 91.96% |
| 7  | facebook.net          | 238 / 56.00%  | 233 / 55.08% |
| 8  | googlesyndication.com | 161 / 37.88%  | 165 / 39.01% |
| 9  | adnxs.com             | 141 / 33.18%  | 139 / 32.86% |
| 10 | bing.com              | 141 / 33.18%  | 144 / 34.04% |

### 5.3 Trace of third-party tracking requests

Here, we investigate the initiator of third-party tracking requests based on the initiator data of each tracking request that *SSOLo* collected. We extract *initiator\_Url* is the URL of the initiator, *call\_Stack* is the function call order of the last function - *initiator\_Function* that sends the tracking request. From that, with each tracking request, we trace back to the origin of these requests. Overall, we find out that there are two types of initiators to send a tracking request are *direct from the first-party* (e.g., *example.com* → *a.track.com*) and *indirect through intermediary* (e.g., *example.com* → *script.analysis.com* → *b.track.com*).

Following that method, we build a tracking mapper for the SSO logins process; we summarize tracking flows at Table 7. Particularly, we detect that 415 first parties send directly more than 80% of third-party web tracking, in which 81.75% (784 out of 959) for Google SSO login and 82.08% (774 out of 943) for Facebook SSO login. In case tracking requests are sent indirectly from first-party through intermediaries, we detect 425 first parties through 637 third-party act as intermediaries then send to 686 third-party tracking (71.53%)

**Table 6: The list of third-party web tracking domains is available only at a single SSO.**

| Only available with Google IdP |                 | Only available with Facebook IdP |                 |
|--------------------------------|-----------------|----------------------------------|-----------------|
| third-party                    | # first parties | third-party                      | # first-parties |
| thebrighttag.com               | 2               | bannerflow.net                   | 4               |
| ipredictive.com                | 2               | adrta.com                        | 2               |
| oracleinfinity.io              | 1               | serving-sys.com                  | 2               |
| affirm.com                     | 1               | cookieless-data.com              | 2               |
| optimonk.com                   | 1               | getsmartcontent.com              | 1               |
| capterra.com                   | 1               | baidu.com                        | 1               |
| webofknowledge.com             | 1               | ayads.co                         | 1               |
| jst.ai                         | 1               | ipinfo.io                        | 1               |
| caffeine.tv                    | 1               | sailthru.com                     | 1               |
| srvsynd.com                    | 1               | px-client.net                    | 1               |

for Google SSO; and 419 first parties through 624 third-party act as intermediaries then send to 674 third-party tracking (71.47%) for Facebook SSO. It is evident that third-party tracking is not only sent directly from the first party but is also integrated and cleverly concealed through a network of third-party intermediaries. Interestingly, the number of tracking requests significantly increases when they pass through these intermediaries, surpassing the number of tracking requests sent directly. The integration of multiple intermediaries forms a complex web that amplifies the volume of tracking requests. This process also presents a challenge for tracking filtering tools.

### 5.4 Effect of SSO logins on online tracking

In this section, we analyze the impact of logging in with and without SSO on third-party tracking.

We provide a comparison of the number of third-party tracking domains before and after logging in using SSO in Figure 4. The dashed line represents the diagonal, indicating that the number of third-party tracking domains before and after SSO login on the same site is equal. Plots above the diagonal indicate an increase in the number of third-party tracking domains after logging into SSO, and plots below the dashed line indicate a decrease. Overall, we observe that most of the plots are concentrated around the diagonal and within the range of 0 to 60 on both horizontal and vertical axes, regardless of whether the user logged in with Google (see Figure 4a) or Facebook (see Figure 4b). This indicates a minimal change in third-party tracking domains before and after signing in with SSO for most sites. However, there are several points above the diagonal, suggesting an increase in the number of third-party tracking domains after SSO login on the same site. Furthermore, there was a significant increase in requests sent to third-party domains after logging in with SSO. Notably, when logging in with Google, the density of plots above the diagonal is higher compared to logging in with Facebook. This indicates that logging in with Google involves a greater number of third-party tracking compared to Facebook.

In addition, we provide a summary based on the top 20 highlighted third-party tracking domains that are only visible on the target website after logging in with SSO (Table 8). Upon investigating the tracking requests from these domains, we have discovered that the requests are generated by factors such as the site's HTML source code, embedded JavaScript snippets, or user navigation. When examining the JavaScript functions responsible for generating tracking requests, we find that most of them are generated by an *anonymous* function. This function has no name and is used to directly call blocks of code without reusing them. Additionally, some

**Table 7: Breakdown flows of third-party tracking.**

| Initiator   | From sender to receiver |                       |
|---|-------------------------|-----------------------|
|   | G                       | F                     |
| Direct from 1st-party<br>by # tracking requests           | 415→784<br>61,102       | 415→774<br>52,390     |
| Indirect through intermediaries<br>by # tracking requests | 425→637→686<br>62,279   | 419→624→674<br>53,100 |

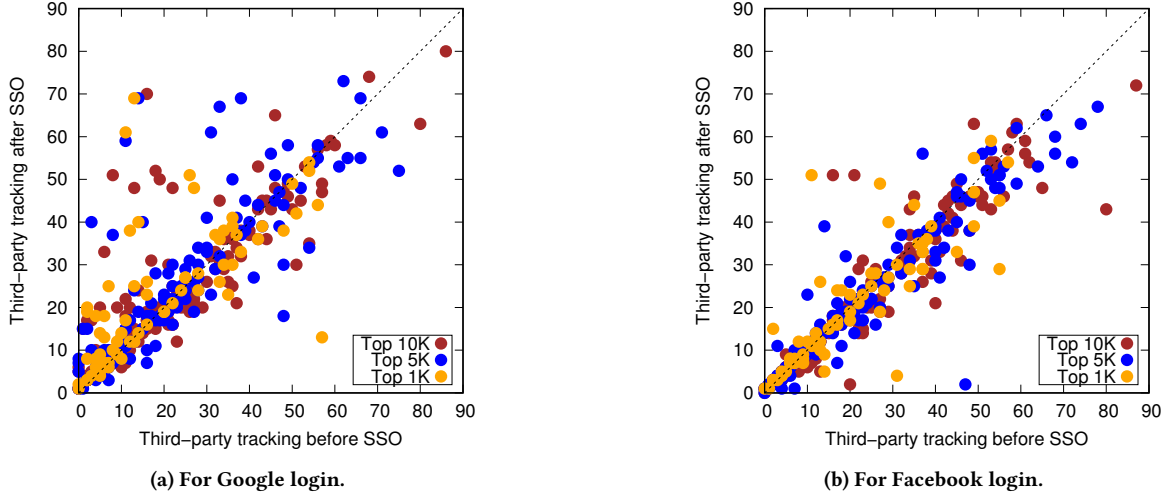


Figure 4: Compare third-party tracking domain before and after SSO login.

Table 8: Breakdown of highlighted third-party tracking only available on the same website after executing SSO login process.

| For Google login      |               |                             |                                 | For Facebook login    |               |                             |  |
|-----------------------|---------------|-----------------------------|---------------------------------|-----------------------|---------------|-----------------------------|--|
| Third-party           | First parties | Initiator type <sup>1</sup> | Initiator function <sup>2</sup> | Third-party           | First parties | Initiator type <sup>1</sup> | Initiator function <sup>2</sup>            |
| media6degrees.com     | 5             | [parser, other]             | n/a                             | colossussp.com        | 3             | [parser, other]             | n/a  |
| moatpixel.com         | 4             | [parser, other]             | n/a                             | nrich.ai              | 3             | [parser, other]             | n/a  |
| dstillery.com         | 3             | script                      | [A, anonymous]                  | brand-display.com     | 2             | other                       | n/a  |
| connexity.net         | 3             | [script, parser, other]     | Manager.sendExternalTracking    | truoptik.com          | 2             | [parser, other]             | n/a  |
| onaudience.com        | 3             | [script, parser, other]     | Manager.sendExternalTracking    | serving-sys.com       | 2             | [script, other]             | anonymous                                  |
| truoptik.com          | 2             | [parser, other]             | n/a                             | adxpremium.services   | 2             | script                      | anonymous                                  |
| ipredictive.com       | 2             | other                       | n/a                             | adxbid.info           | 2             | script                      | [ve, q]                                    |
| thebrighttag.com      | 2             | parser                      | n/a                             | cpmstar.com           | 2             | other                       | n/a  |
| adxpremium.services   | 2             | script                      | anonymous                       | infolinks.com         | 2             | [script, parser]            | [nodeScriptReplace, eval, set, loadScript] |
| adxbid.info           | 2             | script                      | [ve, q]                         | px-client.net         | 1             | script                      | send                                       |
| iteratehq.com         | 1             | script                      | [e, 329]                        | tokbox.com            | 1             | [script, other]             | anonymous                                  |
| fireworkanalytics.com | 1             | script                      | get                             | expedia.com           | 1             | script                      | anonymous                                  |
| extole.io             | 1             | script                      | loader                          | sailthru.com          | 1             | other                       | n/a  |
| tokbox.com            | 1             | [script, other]             | anonymous                       | getsmartcontent.com   | 1             | script                      | [anonymous, X]                             |
| webofknowledge.com    | 1             | [parser, script]            | o                               | yieldlab.net          | 1             | script                      | iaw.activateYieldLabUserMatching           |
| captcha-display.com   | 1             | script                      | anonymous                       | kiosked.com           | 1             | script                      | lb   |
| pghub.io              | 1             | script                      | anonymous                       | seadform.net          | 1             | parser                      | n/a  |
| capterra.com          | 1             | script                      | anonymous                       | adnetmedia.lt         | 1             | parser                      | n/a  |
| srvsynd.com           | 1             | script                      | [value, xFn, anonymous, t.uA]   | ipinfo.io             | 1             | script                      | nrWrapper                                  |
| webvisor.org          | 1             | other                       | n/a                             | fireworkanalytics.com | 1             | script                      | get  |

<sup>1</sup> Initiator of request {*parser*: HTML source, *script*: JavaScript, *other*: user navigate}; <sup>2</sup> The JavaScript function that calls requests, is available if the initiator type is *script*.

functions have names that suggest their involvement in sending tracking data, such as *sendExternalTracking*, *send*, *get*, and *loader*. Our research findings can help in the development of solutions aimed at preventing and securing user data in future projects.

## 6 CONCLUSION

In this paper, we developed the *SSOLogin* framework for SSO login privacy measurement. Through the comprehensive analysis, we demonstrated the effectiveness of our framework in the detection and registration form completion during the SSO login process. We then use this framework to characterize websites containing SSO logins. The results show that 1,420 sites in the Tranco top 10K sites in July 2023 contain SSO logins (Google/Facebook) primarily on websites in the United States and the United Kingdom, within the Information Technology and News/Media website categories.

Finally, we conducted privacy analyses on the login traces. We revealed that most SSO login processes come with online third-party tracking, which may pose risks to user privacy. Moreover, these third parties have established mechanisms to collect and utilize user data from multiple sources, indicating a potential privacy concern. Despite the convenience of logging in with an SSO account, it is important to be aware of potential implications for user privacy.

**Software and dataset availability:** Due to the ethical consideration of releasing automated account registration tools to the public, *SSOLogin* and the crawled dataset will be available from the authors on request for research purposes.

**Acknowledgments:** The authors thank the NII internship program, anonymous reviewers, and Johan Mazel for their valuable feedback that improved our paper.



## REFERENCES

- [1] 2009. Federal Trade Commission. Self-regulatory principles for online behavioral advertising: Tracking, targeting, and technology. (2009). Retrieved December 30, 2022 from <https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-staff-report-self-regulatory-principles-online-behavioral-advertising/p085400behavareport.pdf>
- [2] 2016. Adblockparser limitations. (2016). Retrieved June 20, 2023 from <https://github.com/scraperhub/adblockparser#limitations>
- [3] 2016. Python parser for Adblock Plus filters. (2016). Retrieved March 22, 2023 from <https://github.com/scraperhub/adblockparser>
- [4] 2022. Free IP Geolocation API. (2022). Retrieved August 30, 2022 from <https://freegeoip.app/>
- [5] 2022. IP Geolocation API. (2022). Retrieved August 30, 2022 from <https://ip-api.com/>
- [6] 2023. EasyList. (2023). Retrieved March 30, 2023 from <https://easylist.to/easylist/easylist.txt>
- [7] 2023. EasyPrivacy. (2023). Retrieved March 30, 2023 from <https://easylist.to/easylist/easyprivacy.txt>
- [8] 2023. FortiGuard Web Filtering. (2023). Retrieved January 30, 2023 from <https://fortiguard.com/webfilter>
- [9] 2023. MaxMind GeoIP2 Python API. (2023). Retrieved August 30, 2022 from <https://dev.maxmind.com/geoip/geoip2/geolite2/>
- [10] Rebecca Balebako, Pedro Leon, Richard Shay, Blase Ur, Yang Wang, and L Cranor. 2012. Measuring the effectiveness of privacy tools for limiting behavioral advertising. In *W2SP'12-SP*.
- [11] Lujo Bauer, Cristian Bravo-Lillo, Elli Fragkaki, and William Melicher. 2013. A comparison of users' perceptions of and willingness to use Google, Facebook, and Google+ single-sign-on functionality. In *Proceedings of the 2013 ACM workshop on Digital identity management*. 25–36.
- [12] France Belanger, Janine S Hiller, and Wanda J Smith. 2002. Trustworthiness in electronic commerce: the role of privacy, security, and site attributes. *The journal of strategic Information Systems* 11, 3-4 (2002), 245–270.
- [13] Nataliia Bielova. 2017. Web tracking technologies and protection mechanisms. In *Proceedings of ACM SIGSAC CCS*. 2607–2609.
- [14] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. 2015. I always feel like somebody's watching me: measuring online behavioural advertising. In *Proceedings of ACM CoNEXT*. 1–13.
- [15] Manolis Chatzimpyrros, Konstantinos Solomos, and Sotiris Ioannidis. 2019. You shall not register! detecting privacy leaks across registration forms. In *Computer Security*. Springer, 91–104.
- [16] Ha Dao and Kensuke Fukuda. 2021. Alternative to third-party cookies: investigating persistent PII leakage-based web tracking. In *Proceedings of ACM CoNEXT*. 223–229.
- [17] Steven Englehardt et al. 2018. Automated discovery of privacy violations on the web. (2018).
- [18] Louis Jannett, Vladislav Mladenov, Christian Mainka, and Jörg Schwenk. 2022. DISTINCT: Identity Theft using In-Browser Communications in Dual-Window Single Sign-On. In *Proceedings of the CCS*. 1553–1567.
- [19] Nikhil Jha, Martino Trevisan, Luca Vassio, and Marco Mellia. 2022. The Internet with privacy policies: Measuring the Web upon consent. *ACM Transactions on the Web (TWEB)* 16, 3 (2022), 1–24.
- [20] Chung Hun Lee and David A Cranage. 2011. Personalisation–privacy paradox: The effects of personalisation and privacy assurance on customer responses to travel Web sites. *Tourism Management* 32, 5 (2011), 987–994.
- [21] Wanpeng Li, Chris J Mitchell, and Thomas Chen. 2019. Oauthguard: Protecting user security and privacy with oauth 2.0 and openid connect. In *Proceedings of the SSR*. 35–44.
- [22] Aleecia M McDonald and Lorrie Faith Cranor. 2010. Americans' attitudes about internet behavioral advertising practices. In *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*. 63–72.
- [23] Srivathsan G Morkonda, Sonia Chiasson, and Paul C van Oorschot. 2021. Empirical analysis and privacy implications in OAuth-based single sign-on systems. In *Proceedings of the WPES*. 195–208.
- [24] Paul A Pavlou. 2011. State of the information privacy literature: Where are we now and where should we go? *MIS quarterly* (2011), 977–988.
- [25] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Koryński, and Wouter Joosen. 2019. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Proceedings of NDSS*.
- [26] Chrome DevTools Protocol. [n. d.]. Chrome DevTools Protocol. ([n. d.]). Retrieved Feb 8, 2023 from <https://chromedevtools.github.io/devtools-protocol/tot/Network>
- [27] Ali Rasaii, Shivani Singh, Devashish Gosain, and Oliver Gasser. 2023. Exploring the Cookieverse: A Multi-Perspective Analysis of Web Cookies. In *Proceedings in PAM*. Springer, 623–651.
- [28] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. 2019. TALON: an automated framework for cross-device tracking detection. In *Proceedings of RAID*. 227–241.
- [29] 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*. 1–15.
- [30] Zachary Weinberg, Shinyoung Cho, Nicolas Christin, Vyas Sekar, and Phillipa Gill. 2018. How to catch when proxies lie: Verifying the physical locations of network proxies with active geolocation. In *Proceedings of ACM IMC*. 203–217.
- [31] Maximilian Westers, Tobias Wich, Louis Jannett, Vladislav Mladenov, Christian Mainka, and Andreas Mayer. 2023. SSO-Monitor: Fully-Automatic Large-Scale Landscape, Security, and Privacy Analyses of Single Sign-On in the Wild. *arXiv preprint arXiv:2302.01024* (2023).
- [32] Yuchen Zhou and David Evans. 2014. SSOScan: Automated testing of web applications for Single Sign-On vulnerabilities. In *Proceedings of the USENIX Security*. 495–510.