The Prevalence of Single Sign-On on the Web: Towards the Next Generation of Web Content Measurement

Calvin Ardi
USC/ISI

Matt Calder
Meta / Columbia University

ABSTRACT

Much of the content and structure of the Web remains inaccessible to evaluate at scale because it is gated by user authentication. This limitation restricts researchers to examining only a superficial layer of a website: the landing page or public, search-indexable pages. Since it is infeasible to create individual accounts across thousands of webpages, we examine the prevalence of Single Sign-On (SSO) on the web to explore the feasibility of using a few accounts to authenticate to many sites. We find that 58% of the top 10K websites with logins are accessible with popular 3rd-party SSO providers, such as Google, Facebook, and Apple, indicating that leveraging SSO offers a scalable solution to access a large volume of user-gated content.

CCS CONCEPTS

• **Networks** → **Network measurement**; • **Information systems** → **World Wide Web**; • **Security and privacy** → **Authentication**.

KEYWORDS

Single Sign-On, Web authentication, Top lists, Web measurement

ACM Reference Format:
Calvin Ardi and Matt Calder. 2023. The Prevalence of Single Sign-On on the Web: Towards the Next Generation of Web Content Measurement. In *Proceedings of the 2023 ACM Internet Measurement Conference (IMC '23), October 24–26, 2023, Montreal, QC, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3618257.3624841

1 INTRODUCTION

The web is a critical resource for disseminating information on the Internet and numerous studies have characterized network performance, user experience, complexity, and hosting infrastructure. Previous work has largely relied on profiling the landing page, the default page a user views when navigating to a site (i.e., `index.htm`) [4, 20, 21]. Yet, there has been consensus in measurement community for some time that landing pages are not representative [7].

Hispar [7] made progress beyond landing pages with a technique of using web search to identify the top "internal pages" of a website. Their work found significant differences in structure and performance between landing and internal pages. However, internal pages also have limited representativeness of the pages that users actually interact with.

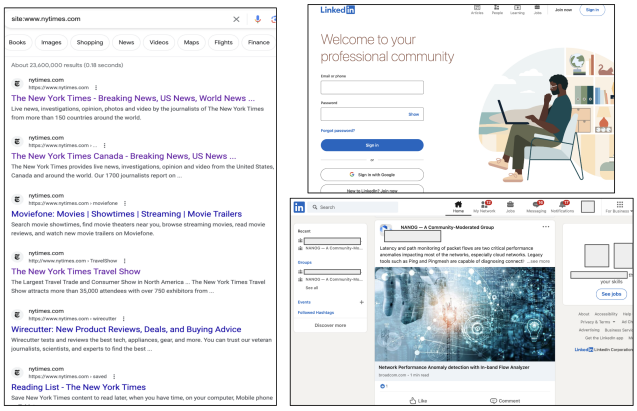

Figure 1: (Left) The top internal pages from the New York Times website from Google search used by Hispar to identify internal pages. (Right) The top shows LinkedIn landing page when logged out. The bottom shows the landing page when a user is logged-in.

Search Restrictions. The first issue is that search engines only find the pages they are allowed by a site, which are not necessarily representative. Figure 1 (left) shows the top internal pages for the New York Times. The results are not popular news stories, but come from the `Allow` paths in the `robots.txt` file [22].

Logged-in user content. Page content and structure is often different depending on whether a user is logged-in. Figure 1 (right) shows the landing pages of LinkedIn when a user is logged-out (top) and logged-in (bottom). When a user is not logged-in, the landing page often *is* the login page.

When most users visit `www.linkedin.com`, the landing page is going to display *personalized* content as the user will already be authenticated. Personalized content, such as news, ads, and recommendations, impact webpage performance because they are often dynamically generated in a datacenter in contrast to the CDN edge serving static content near the user. Realistic web performance is important to measure for both industry due to impact on revenue [19, 29] and researchers' work to improve browsing quality of experience.

Previous work observed that logged-in and non-logged-in users face different privacy concerns, but conducting that study required *manual account creation* for 345 websites from the Alexa Top 500 [15]. Manual account creation does not scale to the hundreds of thousands to millions of sites that modern web measurement studies require.

As a solution to this challenge, we propose leveraging popular SSO providers to access many sites with few accounts. Our goal is to determine the prevalence of SSO Identity Providers (IdPs) on the web to understand how much user-gated content can be accessed.

IMC '23, October 24–26, 2023, Montreal, QC, Canada

ACM ISBN 979-8-4007-0382-9/23/10…$15.00
https://doi.org/10.1145/3618257.3624841

Our approach uses browser automation to visit the Chrome UX Report (CrUX) [27] top websites and initiate sign in to capture the Open Authorization (OAuth) providers available.

We present the following contributions:

- We evaluate two novel measurement techniques that enable large scale discovery of SSO-enabled sites on the web and present the first large scale study of SSO adoption.
- Of the top 10K sites, we find that 51% have a login that may gate access, and more than half of those offer 3rd-party SSO login: 30% of all sites.
- The most popular SSO providers are Google, Facebook, and Apple. These three enable sign-in for 47% of all sites with login and 24% of the top 10K sites.

While our results show promise for large scale measurement of user-gated content on the Web, there still remain a number of challenges which we discuss in §6. All code, data, and artifacts can be found at https://webmeasurements.org.

2 BACKGROUND AND RELATED WORK

Background: SSO is a process that enables users to log in with a single ID on other websites that support it. For example, a user can use their existing Google account (`google.com`) to log in to Stack Overflow (`stackoverflow.com`) or the New York Times (`nytimes.com`) without having to create another set of account credentials. Figure 2 shows an example of logging in to Stack Overflow using a Google account.

As a simplified generalization, the SSO system contains two components: the IdP, which handles authentication[1], and the Service Provider (SP), which provides a web application or access to protected resources. When a user visits Stack Overflow, an SP, they may login by authenticating with a supported 3rd-party IdP (Google, GitHub, or Facebook) or 1st-party authentication (i.e., creating an account directly with Stack Overflow). In this paper, we consider only public and freely available IdPs as shown in Table 1, and do not consider enterprise SSO.

Related Work: There is a large volume of work dedicated to analyzing the structure, performance, and serving infrastructure of the web [4, 9–12, 16, 18, 31]. Prior work examined the privacy differences between logged-in and not logged-in users for 345 websites in the Alexa top sites by manually creating accounts on each one. They found logged-in users were subjected to more ads and more privacy leaks [15]. Recent privacy-focused work found wide-spread access of user data by four SSO IdPs in the Alexa Top 500. Xin et al. investigated how often 1st-party text passwords are passed through third-party CDNs [33]. Our work is focused on an approach to scale logged-in pages to thousands of popular sites to better understand the content, structure, and serving infrastructures.

3 METHODOLOGY

Next, we present our approach for measuring the prevalence of SSO IdPs on the web. We first describe the common login patterns on the web that we use to automatically identify a login page, and then describe how our Crawler uses these patterns to identify IdPs.

[1]Authentication and authorization is typically handled with OAuth [14] or SAML. While OAuth does not use IdP and SP terminology, using consistent terms is helpful in our discussion. The OAuth corollaries for an IdP is the Authorization and Resource Server, and for an SP is the Client Application.

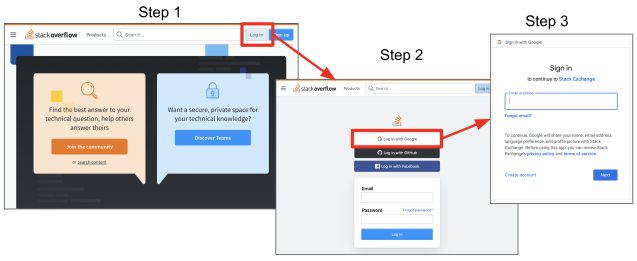

Figure 2: Login flow for Stack Overflow (`stackoverflow.com`), which supports SSO with multiple IdPs: Google, GitHub, and Facebook.

Table 1: Attributes of SSO-Supported Websites

Attribute	Common Values
Login Text	Login, Log in, Sign in, Account, or "My —" (e.g., "My Account" or "my service")
SSO Providers	Amazon, Apple, GitHub, Google, Facebook, LinkedIn, Microsoft, Twitter, Yahoo
SSO Logos	
SSO Text	Sign up with, Sign in with, Continue with, Log in with, Login with, Register with

3.1 Web Login Patterns

Despite the diversity of website design across millions of sites, we have observed some well established patterns in user login interfaces. Figure 2 shows the authentication flow for the popular developer site Stack Overflow which uses the phrase "Log in" for the button (Step 1) and "Log in with", accompanied by company logos, for 3 SSO providers (Step 2): Google, GitHub, and Facebook. We manually inspect 200 CrUX pages with SSO providers and identify the mostly commonly used patterns for both login and SSO buttons seen in Table 1.

3.2 Crawling the Landing and Login Pages

To collect data on SSO providers on a website, we built a Crawler application. After a page completes loading, the Crawler uses a regular expression to find a login button by searching the Document Object Model (DOM) for the common Login Text patterns as shown in Table 1. If found, the Crawler clicks the login button to reach the login page. The Crawler collects screenshots of the landing and login pages, a log of SSO and 1st-party authentication options discovered, and the HTTP transaction log (HAR format).

The Crawler uses the Microsoft Playwright [24] browser automation framework and stable release channel of Google Chrome. We use a plugin to auto-accept cookie banners but not to circumvent bot-detection measures.

3.3 Identifying SSO IdPs on Login Pages

Once our Crawler navigates to the login page, we need to identify how a user can authenticate: through 1st-party authentication, typically containing user/password form fields, or 3rd-party SSO IdPs, typically with a link or button that opens a new window when

Figure 3: Debug view of our logo detection on the Stack Overflow login. Color-coded outlines are drawn around detected SSO IdPs.

clicked. We have two methods for discovering SSO IdPs on websites: DOM-based Inference, which uses a set of XML Path Language (XPath) expressions and heuristics, and logo detection, which uses a template matching technique common in image processing.

3.3.1 DOM-based Inference. When the Crawler reaches a login page, we search the DOM of all website frames for links and buttons with text elements containing common SSO login patterns. We use a precomputed regular expression consisting of all combinations of *SSO Text* and *SSO Providers* from Table 1 (e.g., "Continue with Apple", "Continue with Google") in an XPath selector. The Crawler logs all matches for later analysis.

3.3.2 Logo Detection. Our second method for detecting SSO IdPs is to detect company logos (Table 1) that commonly appear on the login page. We detect logos using template matching which takes a small logo image as the "template" and then scans a larger image to check if the template exists within it [32]. While we initially tried more sophisticated feature-detection approaches such as SURF [8], YOLOv2 [25], and Google's Cloud Vision API [13], we found that they performed poorly due to the small feature space of the logos and the relatively tiny portion of the website that the logo occupies. After manually cropping site screenshots to where the logo was a much larger portion of the overall image, we saw improvement but chose to stick with a simple approach.

Logo detection uses template matching from OpenCV [23]. Because template matching does not support multiple-scale detection, we use the common approach [3] of rescaling a single template to 10 different sizes to capture variation across websites.

We manually collected multiple logo templates from login pages of 100 pages to capture variations. The Google logo is quite consistent, while Twitter and Apple each have light and dark variations (,). Facebook has the largest number of variations with different light and dark schemes mixed with square and round backgrounds, and centered and offset lower-case "f" (,).

The input for logo detection is the set of templates and a collection of login page screenshots taken by the Crawler. For each image, the logo detection iterates through all templates for each SSO IdP. When logo detection detects a match with at least 90% probability, it flags that IdP as seen, and the logo detection continues to the next IdP (Figure 3). While this brute force approach is slow, it parallelizes easily and took approximately 45 min to run on 1000 sites using 7 CPU cores.

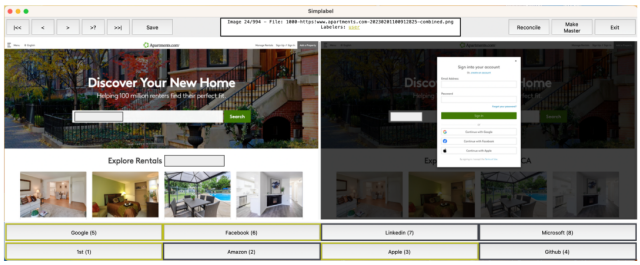

Figure 4: Simplabel, an image labeling software, extended to support multiple labels and side-by-side images.

3.4 Limitations

Our data collection and results are a lower bound on the amount of opportunity for using SSO to automate large-scale login for web measurements for several reasons. Our approach relies on popular design practices, but the web is complicated and dynamic, with a long tail distribution in variation: our Crawler is unable to crawl certain sites due to some common problems (§6).

Both our inference approaches have limitations. DOM-based Inference is closely tied to language-specific expressions which (currently) must be manually curated. It is possible that login pages of non-English websites do not share the common patterns that we leverage. Additionally, our web login patterns are not exhaustive and may not include patterns found in unpopular content or non-English sites.

Logo detection is language agnostic but is susceptible to false positives from ads or content related to an IdP but not directly related to SSO functionality (§4.2). Also, we likely exclude regionally popular SSO providers. Finally, our measurements are biased towards popular content. It is possible that our results are not representative of less popular sites.

4 VALIDATION

To validate our ability to identify SSO IdPs on websites, we first build and manually label a ground truth dataset. We then measure the performance (precision and recall) of our DOM-based inference and logo image detection techniques.

4.1 Building a Ground Truth Dataset

We first use the Crawler to profile the Top 1K (50% of Chrome page loads [26]) CrUX U.S. websites from the February 2023 list, which we collect from BigQuery [1].

To assist with ground truth labeling of our top 1K crawl, we extend and use an open-source tool, Simplable [2], to support multiple labels and side-by-side view of the landing page and login page for each site (Figure 4). The labeling task consists of (1) identifying if there is a login button, (2) if the Crawler clicked it successfully, and (3) detecting what 1st-party authentication and 3rd-party SSO IdPs are present.

Table 2 shows the results of our ground truth labeling. Of the Top 1K sites crawled, 994 were responsive. We label around 28% as

Table 2: Crawler Performance and IdPs of the Top 1K

Description	%	%*	%*	#
Total	100.0			994
Broken	27.7			275
Blocked	8.0			80
Successful	64.4	100.0		640
3rd-party SSO IdP		31.6		202
Google			89.6	181
Facebook			60.4	122
Apple			48.0	97
Other			18.3	37
Microsoft			5.9	12
Twitter			5.9	12
Amazon			3.5	7
Linkedin			2.5	5
Yahoo			2.0	4
Github			0.5	1
1st-party Login		77.7		497
No Login		20.8		133

* Total is over 100% as a website can support many IdPs.

broken because they contained a login button that the Crawler failed to detect or click correctly (§6). 8% of sites which are labeled *blocked* used some kind of bot-detection service preventing our Crawler from loading the page. While there are methods to circumvent this, we opted not to for ethical reasons (Appendix B). Our Crawler is successful for 64% of sites by correctly navigating to the login page or finding that there is no login. We find 31.6% of successfully crawled sites contain at least one SSO IdP, while 77.7% contain 1st-party authentication.

The presence of broken sites causes us to underestimate the number of pages with logins and SSO. Since the number of successful pages is 640, we believe this is large enough sample that our results are not significantly impacted.

4.2 How does our Inference Perform?

Table 3 summarizes the results (precision, recall, F_1 score) of using our two inference techniques, both independently and together.

Using DOM-based Inference for all 10 providers is very precise (little to no false positives), with scores between 0.97–1.00. However, recall varies greatly between providers as this technique frequently misses (many false negatives) IdP, especially LinkedIn, Microsoft, and Yahoo. One of the challenges with DOM-based Inference is its reliance on expecting certain text patterns. For example, it may miss an SSO IdP if the SSO button is a custom-drawn logo without any accompanying text.

Logo detection has good precision for popular IdPs, but performs poorly for Amazon, Twitter, and Microsoft, leading to a high false positive rate for these particular providers. We note that the population of sites that use less popular IdPs is an order of magnitude smaller than the most popular IdPs (Google, Facebook, Apple). The majority of false positives in logo detection come from social media links embedded on the website (e.g., links to a Twitter or Facebook profile), the Apple App Store logo (e.g., a link to the entity's mobile application), or ads for Amazon and Microsoft products. We present an example visualization in our logo detection application demonstrating a false positive in Appendix A.

Table 3: Performance of Finding IdPs in Top 1K

IdP	DOM-based			Logo Detection			Combined		
	P	R	F_1	P	R	F_1	P	R	F_1
Google	0.98	0.68	0.80	0.99	0.93	0.96	0.97	0.97	0.97
Facebook	0.99	0.73	0.84	0.76	0.80	0.78	0.78	0.91	0.84
Apple	0.97	0.75	0.85	0.80	0.94	0.86	0.80	0.98	0.88
Microsoft	1.00	0.42	0.59	0.39	0.58	0.47	0.39	0.58	0.47
Twitter	1.00	0.45	0.63	0.19	1.00	0.31	0.19	1.00	0.31
Amazon	1.00	1.00	1.00	0.38	0.86	0.52	0.41	1.00	0.58
LinkedIn	1.00	0.20	0.33	—	—	—	1.00	0.20	0.33
Yahoo	1.00	0.25	0.40	1.00	0.75	0.86	1.00	1.00	1.00
GitHub	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1st-party	0.99	0.61	0.76	—	—	—	0.99	0.61	0.76

P = Precision, R = Recall

Table 4: 1st-party vs. SSO Logins on Websites

Description	Top 1K		Top 10K	
	%	#	%	#
SSO or 1st-party	100.0	507	100.0	4743
1st-party only	60.2	305	42.2	2001
SSO and 1st-party	37.9	192	23.3	1107
SSO only	2.0	10	34.5	1635
No Login, Broken, or Blocked		488		4530

Recall in logo detection is better overall compared to DOM-based Inference, which results in its ability to better find IdPs when they are actually present.

Finally, we combine the results of DOM-based Inference and logo detection together by doing a binary `OR` on the results of each technique. With this combination, we generally find an increase in recall (less false negatives) with a decrease in precision (more false positives). We find this trade-off acceptable, and we note that the scores for less popular IdPs are skewed due to the much lower population size.

5 PREVALENCE OF SSO ON THE WEB

We next evaluate and discuss the prevalence of 1st- and 3rd-party SSO authentication on supported websites. We frame our discussion around the top 1K and top 10K websites, which make up a huge portion of popular web content. Ruth et al. found that the top 1K sites make up 50% of all page loads and the top 10K constitutes 70% (~40% were the top 100, but the smallest bin in the public CrUX list is 1K) [26].

5.1 How Many Sites Support SSO?

We first measure how many sites have a login mechanism, and whether they support 3rd-party SSO authentication.

SSO is a popular way to extend the login capabilities of a website: 51% (4743) of sites in the Top 10K and 51% (507) in the Top 1K have a login function—we denote these subsets as Top $10K_L$ and Top $1K_L$, respectively. 57.8% (2742) of these 4743 sites support authentication with 3rd-party SSO IdPs. Table 4 summarizes our results.

Out of the Top $10K_L$ pages, 1st-party only authentication makes up 42.2% (2001) but accounts for the majority (60.2%, 305) in the Top $1K_L$. As described in §2, we determine 1st- or 3rd-party from the

Table 5: SSO IdPs of Top 10K

Description	%	%*	%*	#
Total	100.0			9273
Login	51.1			4743
3rd-party SSO IdP		57.8		2742
Facebook			45.9	1258
Google			39.8	1092
Apple			36.0	986
Twitter			29.7	815
Amazon			5.7	156
Microsoft			4.9	133
LinkedIn			0.3	9
Yahoo			0.3	9
GitHub			0.3	7
1st-party		65.5		3108
No Login	48.9			4530

* Total is over 100% as a website can support many IdPs.

perspective of the SP's entity: for example, a Google account which provides SSO across all of Google's websites (Gmail, YouTube) is considered 1st-party. Thus, the top 1K sites are more likely to have the need for and resources to control all aspects of their authentication flow. While enabling 3rd-party providers would increase user acquisition, it also ties itself to the 3rd-party's reliability: an outage to the 3rd-party SSO IdP results in loss of use of the SP.

A significant portion of sites (34.5%, 1635) in the Top 10K$_L$ support only 3rd-party SSO authentication. Supporting 3rd-party SSO only can be beneficial as it reduces a website owner's responsibility in properly implementing and securing the account creation and login process, and reduces user friction in signing up for a service. For example, Tailscale [30], a VPN service, does not support 1st-party account creation by design, eliminating the need to store usernames and passwords. The trade-off, however, is loss of some control from SP's perspective as they may not have access to or control of all the user's information and data: this is perhaps why only 2.0% (10) of the Top 1K$_L$ are 3rd-party SSO only.

5.2 What are the Popular SSO Providers?

We next measure the popular SSO Identity Providers (IdPs), and determine whether any particular combination of providers is more prevalent than another.

Facebook is the most popular SSO IdP, supported on 45.9% (1258) of the sites in the Top 10K$_L$, with Google (39.8%, 1092), Apple (36.0%, 986), and Twitter (29.7%, 815) supported on roughly the same number of sites. In the Top 1K$_L$, Google (89.6%, 181), Facebook (60.4%, 122), and Apple (48.0%, 97) are the most prevalent. The remaining IdPs (Amazon, Microsoft, etc.) are supported on less than ~6% in both the Top 10K$_L$ and Top 1K$_L$. Table 5 and Table 2 summarize the results for the Top 10K$_L$ and Top 1K$_L$, respectively.

A key result of measuring the popularity of SSO is that having accounts on three SSO IdPs, Google, Apple, and Facebook, is sufficient to log in to 47.2% (2238) of sites that have authentication (81.6% of sites that support 3rd-party SSO). Because these IdPs use the same underlying OAuth mechanism [14], it is relatively easy to add support for multiple providers: while most sites (56.0%, 1536) support only one IdP, 27.2% (747) support two, and 16.7% (459) support three or more (Table 6). A comprehensive list of SSO IdP combinations can be found in Table 8 and Table 9 in Appendix C.

Table 6: Number of SSO IdPs on Websites

# SSO IdPs	Top 1K$_L$		Top 10K$_L$	
	%	#	%	#
Total	100.0	202	100.0	2742
1	21.8	44	56.0	1536
2	32.7	66	27.2	747
3	35.1	71	14.8	406
4	8.4	17	1.8	48
5	1.5	3	0.2	5
6	0.5	1	—	—

Apple's prominence as an IdP may partly be the result of its guidelines [6] (updated in 2019 [5]) requiring developers to integrate Apple's IdP ("Sign in with Apple") in their apps if they use any other 3rd-party IdP. We leave measuring the growth and prominence of SSOs over time as future work.

5.3 What Type of Sites Support SSO?

Finally, we next look at the categories of sites that support 1st- and 3rd-party SSO authentication in the Top 1K.

We find 31–78% of all websites support login authentication (1st- or 3rd-party), and there is at least 1 website (3–36%) in each of nine categories (except Healthcare) which supports 3rd-party SSO. Table 7 summarizes the results.

29–36% of sites in the Business Service, Informational, Social Network, and News categories support 3rd-party SSO, while only one site in Finance and zero in Healthcare offer the same support. Interestingly, 3 Adult sites support SSO through Google, Twitter, or another Adult site. We leave investigation into what data is collected from the IdP and its privacy implications as future work.

The limited-to-no support for 3rd-party SSO in Finance and Healthcare is unsurprising, given the sensitivity and privacy concerns of the data being stored (e.g., HIPAA in the U.S.). As IdPs implement more user-level controls and protections on user data, users on Finance and Healthcare SPs might benefit from 3rd-party SSO integration. Currently users who wish to extract or use data from these sites on another application (e.g., personal finance aggregators) must often pass their username and password credentials directly to the application, which OAuth was designed to prevent.

6 DISCUSSION AND FUTURE WORK

The goal of our work is to enable more representative web measurements. In this section, we review some of the challenges we encountered with this work, what is left on the table, and future directions to explore.

Challenges with Crawling. We ran into a number of common artifacts which prevented our Crawler from finding the login page, but appear solvable with additional work. These include age-verification prompts from adult websites and sales banners from online retailers, which often must be dismissed before any other interaction is possible. A login pattern which contributes to a large number of broken

Table 7: Website Categories and Supported Logins in Top 1K

Description	Biz. Svc.		Shop		Ent.		Lifestyle		Adult		Info.		News		Finance		Social		Health	
	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#
Total	100.0	279	100.0	176	100.0	129	100.0	125	100.0	78	100.0	62	100.0	61	100.0	40	100.0	27	100.0	17
No Login	31.5	88	69.3	122	45.0	58	56.0	70	67.9	53	58.1	36	42.6	26	35.0	14	22.2	6	52.9	9
Login	68.5	191	30.7	54	55.0	71	44.0	55	32.1	25	41.9	26	57.4	35	65.0	26	77.8	21	47.1	8
1st-party only	38.0	106	21.6	38	34.9	45	26.4	33	28.2	22	12.9	8	21.3	13	62.5	25	44.4	12	47.1	8
SSO, 1st-party	29.4	82	9.1	16	19.4	25	15.2	19	3.8	3	24.2	15	36.1	22	2.5	1	33.3	9	0.0	0
SSO only	1.1	3	0.0	0	0.8	1	2.4	3	0.0	0	4.8	3	0.0	0	0.0	0	0.0	0	0.0	0

Biz. Svc. = Business Service, Ent. = Entertainment, Info. = Informational, Social = Social Networking

pages is when the person icon ( ) represents the login button without any accompanying text labels: to resolve this, we are exploring the use of web accessibility features (e.g., `aria-label` [28]).

What We are Missing. Certain critical infrastructure sites like banking and healthcare do not support popular SSO providers, and for security and privacy reasons. This area, however, remains a blind spot of significance for researchers conducting web measurements: these institutions play an important role in our lives and we should continuously measure the performance of their websites to monitor their accessibility. It is also likely difficult to automate on sites that require a paid subscription to function correctly, such as news outlets or streaming video services.

Security and Privacy. As briefly discussed earlier, SSO provides an undetermined trade-off between usability, security, and privacy for both users of and entities operating SPs. While SSO eases account creation and authentication (user) as well as management and user acquisition (entity), it becomes a potential single point of failure (e.g., service outage or blockage) as well as an attractive target for attackers (e.g., phishing, cyberattacks). Further study is needed to understand and quantify the risk that users and entities (both IdPs and SPs) are exposed to for both using and not using SSO. For example, some finance applications work around a bank's lack of API or SSO/OAuth integration by directly logging in on behalf of the user, presenting a different set of security and privacy issues.

Future Work. SSO makes possible the automated login of many sites with a small number of accounts, but evaluation of a robust system to perform this is future work. Several pitfalls are clear now: how many sites will challenge automated login with CAPTCHA, multi-factor authentication (MFA), or rate-limiting? How will a site treat profiles that have no browsing history, personal details, or preferences? How can we seed realistic profiles to observe meaningful content when logged-in?

7 CONCLUSION

In this work, we have discussed the opportunity to extend the representativeness of web measurements by using logged-in pages and presented two novel measurement techniques that enable large scale discovery of SSO-enabled sites on the web. We presented the first large scale study of SSO adoption on the web, and found that accounts with popular IdPs like Google, Facebook, and Apple enable Single Sign-On for over 58% of of the top 10K pages with logins, or 30% of the top 10K overall. We have also outlined a series of challenges for web measurements based on our preliminary work that will drive the area forward.

ACKNOWLEDGMENTS

We would like to thank Ethan Katz-Bassett for detailed feedback and our shepherd Kyle Schomp and the anonymous reviewers for their constructive comments.

REFERENCES

[1] [n. d.]. Cached Chrome Top Million Websites. https://github.com/zakird/crux-top-lists
[2] [n. d.]. Simplabel. https://github.com/hlgirard/Simplabel
[3] Adrian Rosebrock. [n. d.]. Multi-scale Template Matching using Python and OpenCV. https://pyimagesearch.com/2015/01/26/multi-scale-template-matching-using-python-opencv/
[4] Bernhard Ager, Wolfgang Mühlbauer, Georgios Smaragdakis, and Steve Uhlig. 2011. Web Content Cartography. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference* (Berlin, Germany) *(IMC '11)*. Association for Computing Machinery, New York, NY, USA, 585–600. https://doi.org/10.1145/2068816.2068870
[5] Apple. 2019. New Guidelines for Sign in with Apple. https://developer.apple.com/news/?id=09122019b
[6] Apple. 2023. App Store Review Guidelines. https://developer.apple.com/app-store/review/guidelines/#sign-in-with-apple
[7] Waqar Aqeel, Balakrishnan Chandrasekaran, Anja Feldmann, and Bruce M. Maggs. 2020. On Landing and Internal Web Pages: The Strange Case of Jekyll and Hyde in Web Performance Measurement. In *Proceedings of the ACM Internet Measurement Conference* (Virtual Event, USA) *(IMC '20)*. Association for Computing Machinery, New York, NY, USA, 680–695. https://doi.org/10.1145/3419394.3423626
[8] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. SURF: Speeded Up Robust Features. In *Computer Vision – ECCV 2006*, Aleš Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 404–417.
[9] Michael Butkiewicz, Harsha V. Madhyastha, and Vyas Sekar. 2011. Understanding Website Complexity: Measurements, Metrics, and Implications. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference* (Berlin, Germany) *(IMC '11)*. Association for Computing Machinery, New York, NY, USA, 313–328. https://doi.org/10.1145/2068816.2068846
[10] Michael Butkiewicz, Daimeng Wang, Zhe Wu, Harsha V. Madhyastha, and Vyas Sekar. 2015. Klotski: Reprioritizing Web Content to Improve User Experience on Mobile Devices. In *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*. USENIX Association, Oakland, CA, 439–453. https://www.usenix.org/conference/nsdi15/technical-sessions/presentation/butkiewicz
[11] Trinh Viet Doan, Roland van Rijswijk-Deij, Oliver Hohlfeld, and Vaibhav Bajpai. 2022. An Empirical View on Consolidation of the Web. *ACM Trans. Internet Technol.* 22, 3, Article 70 (feb 2022), 30 pages. https://doi.org/10.1145/3503158
[12] Theresa Enghardt, Thomas Zinner, and Anja Feldmann. 2019. Web Performance Pitfalls. In *Passive and Active Measurement*, David Choffnes and Marinho Barcellos (Eds.). Springer International Publishing, Cham, 286–303.
[13] Google. [n. d.]. Detect logos | Cloud Vision API. https://cloud.google.com/vision/docs/detecting-logos
[14] Dick Hardt. 2012. The OAuth 2.0 Authorization Framework. RFC 6749. https://doi.org/10.17487/RFC6749
[15] Andrew J. Kaizer and Minaxi Gupta. 2016. Characterizing Website Behaviors Across Logged-in and Not-Logged-in Users. In *Proceedings of the 2016 Internet Measurement Conference* (Santa Monica, California, USA) *(IMC '16)*. Association for Computing Machinery, New York, NY, USA, 111–117. https://doi.org/10.1145/2987443.2987450
[16] Conor Kelton, Jihoon Ryoo, Aruna Balasubramanian, and Samir R. Das. 2017. Improving User Perceived Page Load Times Using Gaze. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. USENIX Association,

Boston, MA, 545–559. https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/kelton
[17] Nico Kokonas. 2021. Playwright Stealth. https://gist.github.com/nicoandmee/1ec1b6a07c94f82df41d2496194ef3a6
[18] Zhichun Li, Ming Zhang, Zhaosheng Zhu, Yan Chen, Albert Greenberg, and Yi-Min Wang. 2010. WebProphet: Automating Performance Prediction for Web Services. In *7th USENIX Symposium on Networked Systems Design and Implementation (NSDI 10)*. USENIX Association, San Jose, CA. https://www.usenix.org/conference/nsdi10-0/webprophet-automating-performance-prediction-web-services
[19] Greg Linden. 2006. Make Data Useful. http://sites.google.com/site/glinden/Home/StanfordDataMining.2006-11-28.ppt.
[20] Ravi Netravali, Ameesh Goyal, James Mickens, and Hari Balakrishnan. 2016. Polaris: Faster Page Loads Using Fine-grained Dependency Tracking. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*. USENIX Association, Santa Clara, CA. https://www.usenix.org/conference/nsdi16/technical-sessions/presentation/netravali
[21] Ravi Netravali, Vikram Nathan, James Mickens, and Hari Balakrishnan. 2018. Vesper: Measuring Time-to-Interactivity for Web Pages. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*. USENIX Association, Renton, WA, 217–231. https://www.usenix.org/conference/nsdi18/presentation/netravali-vesper
[22] New York Times. 2023. Internet Archive Snapshot of nytimes.com/robots.txt from 2023/05/19. http://web.archive.org/web/20230519003326/https://www.nytimes.com/robots.txt
[23] OpenCV. [n. d.]. OpenCV: Template Matching. https://docs.opencv.org/3.4/d4/dc6/tutorial_py_template_matching.html
[24] Playwright. 2023. Playwright: Fast and reliable end-to-end testing for modern web apps. https://playwright.dev/
[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 779–788. https://doi.org/10.1109/CVPR.2016.91
[26] Kimberly Ruth, Aurore Fass, Jonathan Azose, Mark Pearson, Emma Thomas, Caitlin Sadowski, and Zakir Durumeric. 2022. A World Wide View of Browsing the World Wide Web. In *Proceedings of the 22nd ACM Internet Measurement Conference* (Nice, France) *(IMC '22)*. Association for Computing Machinery, New York, NY, USA, 317–336. https://doi.org/10.1145/3517745.3561418
[27] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. 2022. Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists. In *Proceedings of the 22nd ACM Internet Measurement Conference* (Nice, France) *(IMC '22)*. Association for Computing Machinery, New York, NY, USA, 374–387. https://doi.org/10.1145/3517745.3561444
[28] W3 Schools. [n. d.]. Accessibility Labels. https://www.w3schools.com/accessibility/accessibility_labels.php
[29] Stoyan Stefanov. 2008. YSlow 2.0. In *CSDN SD2C*.
[30] Tailscale Inc. 2023. Supported SSO identity providers. https://tailscale.com/kb/1013/sso-providers/
[31] Xiao Sophia Wang, Aruna Balasubramanian, Arvind Krishnamurthy, and David Wetherall. 2013. Demystifying Page Load Performance with WProf. In *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*. USENIX Association, Lombard, IL, 473–485. https://www.usenix.org/conference/nsdi13/technical-sessions/presentation/wang_xiao
[32] Wikipedia. [n. d.]. Template Matching. https://en.wikipedia.org/wiki/Template_matching
[33] Rui Xin, Shihan Lin, and Xiaowei Yang. 2023. Quantifying User Password Exposure To Third-Party CDNs. In *Passive and Active Measurement: 24th International Conference, PAM 2023, Virtual Event, March 21–23, 2023, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 652–668. https://doi.org/10.1007/978-3-031-28486-1_27

A LOGO DETECTION

Figure 5 shows an example of login button false positives generated by our logo detection application for Twitter, Facebook, and Apple. The icons for Twitter and Facebook are links to the entity's social networking profiles, and the Apple icon is flagged from its App Store link to the company's mobile application.

B ETHICS

This work uses a programmable browser framework, Microsoft Playwright, to automate the browsing of web pages using the Google Chrome web browser, which may be categorized as bot-like behavior. During testing we noticed that some pages use bot detection services, such as Cloudflare, to which caused a user input prompt or outright denied access. While there are plugins and specialized browsers to circumvent these [17], after some discussion we chose not to use these as they intentionally subvert a measure designed to protect a business or site from abuse. From our ground truth labeling dataset, we estimate this impacts roughly 8% of sites and feel this is a reasonable trade-off.

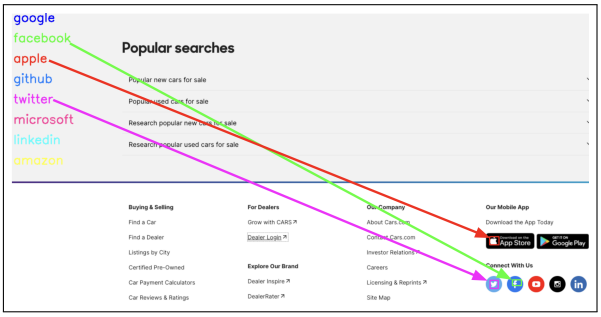

Figure 5: False Positive Result in Logo Detection. Site: `www.cars.com`.

C ADDITIONAL TABLES

Table 8: SSO IdP Combinations in Top 1K$_L$

SSO IdPs	%	#
Total	100.0	201
Apple, Facebook, Google	27.2	55
Google	13.9	28
Facebook, Google	11.4	23
Apple, Google	8.4	17
Google, Other	6.9	14
Facebook	5.4	11
Apple, Facebook, Google, Other	2.5	5
Apple, Facebook, Google, Twitter	2.5	5
Other Combinations	21.8	44

Table 9: SSO IdP Combinations in Top 10K$_L$

SSO IdPs	%	#
Total	100.0	2742
Apple	14.8	407
Google	12.4	339
Twitter	11.8	323
Facebook, Twitter	10.7	294
Facebook	10.7	293
Apple, Facebook, Google	10.0	274
Facebook, Google	7.0	192
Apple, Google	3.9	108
Amazon	3.6	100
Microsoft	2.7	74
Facebook, Google, Twitter	1.6	44
Apple, Facebook, Twitter	1.3	36
Apple, Twitter	1.3	35
Apple, Facebook	1.1	30
Apple, Facebook, Google, Twitter	0.9	25
Other combinations	6.1	168