A Look at the Third-Party Identity Management Landscape

Anna Vapen[†] Niklas Carlsson[†] Anirban Mahanti[‡] Nahid Shahmehri[†]
[†] Linköping University, Sweden, firstname.lastname@liu.se
[‡] NICTA, Australia, anirban.mahanti@nicta.com.au

Abstract

Many websites act as relying parties (RPs) by allowing access to their services via third-party identity providers (IDPs) such as Facebook and Google. Using IDPs simplifies account creation, login, and information sharing across websites. Different websites' use of IDPs can have significant security and privacy implications for the end users. This paper presents an overview of the current state of the third-party identity management landscape. Datasets collected using both manual identification and large-scale crawling are used to answer questions related to which sites act as RPs, which sites are the most successful IDPs, and how different classes of RPs select their IDPs. We also analyze and discuss longitudinal changes in the landscape.

Index Terms

Web authentication, privacy, security, third-party identity provider, identity management, OpenID, OAuth, SSO

I. Introduction

With third-party single sign-on (SSO) services, regular websites such as Huffington Post can allow their users to quickly and easily create personalized user accounts and to login to their site using an account that the user already has with the popular third-party provider. In a third-party SSO scenario, Huffington Post, for example, presents the user with the option of using the user's existing account(s) with Facebook, Twitter, or other IDPs for authentication. Second, having selected Facebook, for example, the user is directed to Facebook for authentication. In this example, Huffington Post is the *relying party* (RP), and Facebook is the third-party *identity provider* (IDP). Together, the RPs and the IDPs make up the third-party identity management landscape.

Today, IDPs are also being used for sharing of personal information across websites [12]. In our example, users are asked if they want to share information from their Facebook account with Huffington Post. Use of the OAuth (authorization) protocol can also enable RPs to post information to the user's IDP account on behalf of the user. Cross-site sharing of personal information can be used to improve end-user service, but these services also introduce privacy and security concerns [1], [12], [13]. The increasing entanglement of RPs and IDPs has implications for end-users as personal information of the user can flow both ways.

The RP-IDP landscape continues to evolve and surprisingly not much is known about which sites are the most frequent IDP users and how different types of RPs select their IDPs. The RPs and their IDP selection play a particularly important role in this emerging landscape. Not only are they affecting the IDP choices provided to end-users, but as the RPs implement the RP-IDP communication, they also directly impact the potential information that may be moved between sites, and hence also the privacy risks faced by the end users. The impact of IDP choices made by the RPs is further illustrated by different IDPs offering different APIs, authentication/authorization protocols, level of security, and information exchange possibilities. In this article, we fill this void by presenting a measurement driven analysis of the RP-IDP landscape.

In this article, we first characterize RPs and other crawled websites. Second, we identify and characterize the most popular IDPs. Third, we characterize which IDPs are selected by different classes. Breaking down the usage within classes of websites, we provide insights to differences in the usage of these services within and across classes, and questions such as why there are so many RPs and so few IDPs, how these IDPs are being used, and why some IDPs are more popular than others. Finally, we use longitudinal data to capture current trends in the third-party identity management landscape.

II. Background and Related Work

Figure 1 shows the evolution of the third-party identity landscape.

Protocol usage: The major protocols for third-party authentication today are OpenID[1] and OAuth[2]. OpenID is an open standard for Web-SSO, while OAuth (who also provides SSO) primarily is an authorization protocol that allows a user to authorize an application on one site to act on behalf of the user on another site. Compared with OpenID, OAuth allows for much richer cross-site sharing and flexible design. The richer information has privacy implications [12] and the flexible design

[1] OpenID Foundation, http://openid.net/, July 2014.
[2] OAuth, http://oauth.net/, July 2014.

This is the authors' version of the work. It is posted here by permission of IEEE for your personal use. Not for redistribution. The definitive version is published in *IEEE Internet Computing* and is available at IEEE Xplore Digital Library.

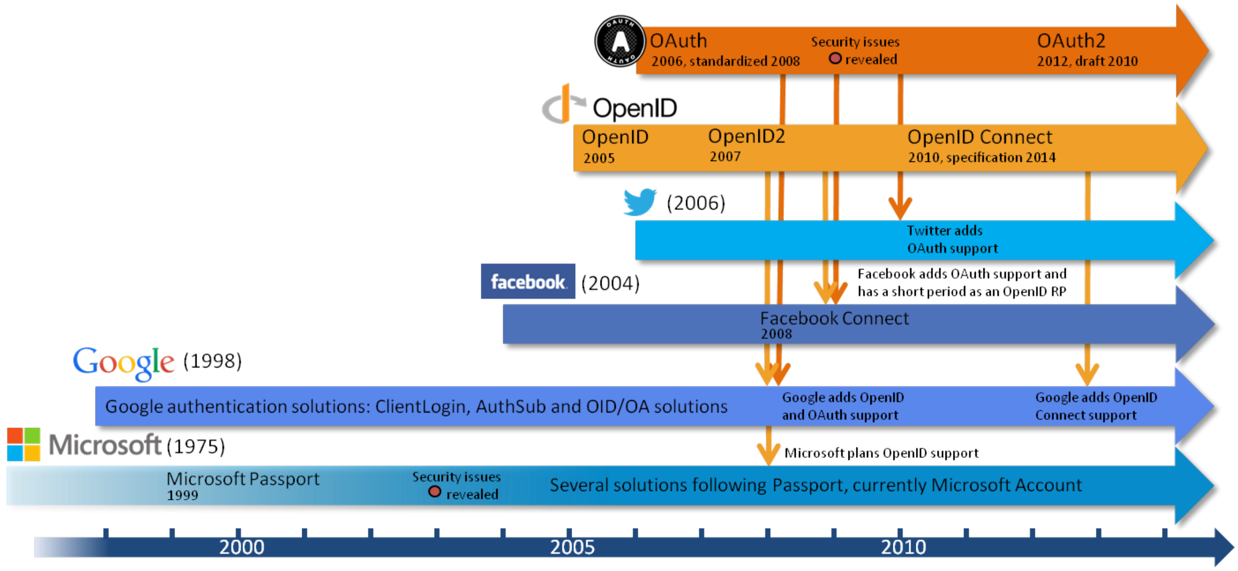


Fig. 1. Timeline of the third-party identity management landscape.

can easily result in security flaws of individual implementations [8]. OpenID Connect combines OpenID with OAuth. SAML SSO is another emerging standard, but is not yet visible in our datasets.

Influential players: Popular websites often act as IDPs, implementing their own solutions based on the popular open protocols discussed above. For example, Twitter incorporated OAuth into its own protocol in 2010 and forced all their RPs to move to this new version. Facebook launched Facebook Connect as its own third-party authentication protocol in 2008. Facebook was also part of the early development of OAuth2, and has used OAuth as part of their protocol since then. Facebook also had a short period during which it acted as an OpenID RP (with Google as IDP).

Google added OpenID support to their own protocol in 2008, shortly followed by OAuth support. Microsoft has a long history of running their own proprietary protocols, and have over the years faced several security problems [3].

Related works: Security weaknesses have been identified in both OAuth and OpenID [8], [10]. However, perhaps the biggest weakness in SSO is the implementation of the protocols themselves [8]. For example, Wang et al. [13] identified eight serious logical flaws which allowed an attacker to sign on as the victim user of a number of high-profile IDPs and RPs. Dhamija et al. [1] describe, discuss, and enumerate seven flaws in federated identity management solutions, each relating to issues such as security, privacy, usability, and economical factors. Other works have considered or characterized specific privacy [4], usability [7], and economic [5] aspects of SSO services.

Pure OpenID-based solutions (which only provide authentication, not authorization as with OAuth), gives very little incentive to act as RP [9]. While several large IDPs have tried OpenID as part of their protocol, many websites instead use OAuth-based IDPs [7], [11]. As shown by our longitudinal dataset, this has resulted in an upswing in IDP usage.

In contrast to the above works, we use a measurement-driven analysis to characterize the observed third-party IDP landscape.

III. Datasets and Relationship Identification

Manual top-200 datasets: We collected datasets by careful manual classification of site-to-site relationships of a large number of potential RP-IDP identified by ourselves, web crawlers, and student volunteers. This dataset tries to identify all relationships involving the top-200 Alexa websites at the time of data collection and have been collected at six different time instances between Apr. 2012 and Feb. 2014.

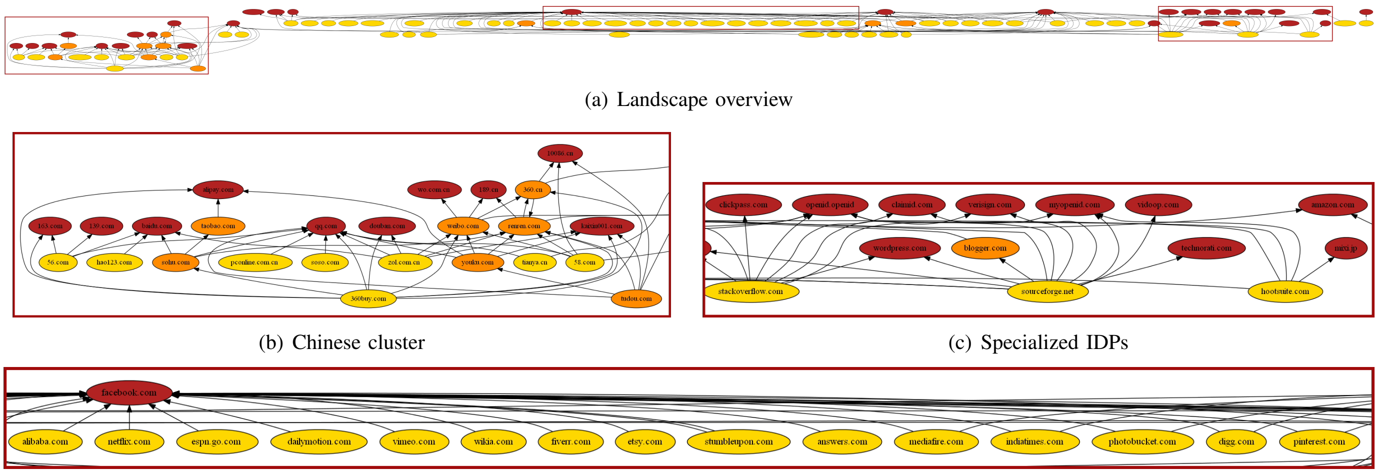
Large-scale dataset: We used popularity-based logarithmic sampling to pick a carefully selected sample set with 35,619 websites from the top-million most popular websites[3]. We used logarithmic sampling because website popularity is known to be long-tailed [2]. We then used our multithreaded Selenium-based[4] crawler [11] to crawl these sites and identify RP-IDP relationships, while acting like a human user that clicks on the available GUI elements, reacts to popups, and takes other GUI-based actions.

While our tool does not find all RP-IDP relationships, our manual validation has shown that the tool is successful in avoiding false positives [11]. Out of the 35,619 sampled websites, we found 1,865 RPs, 50 IDPs, and 3,329 unique RP-IDP relationships.

Complementing information: During our large-scale crawl we also collected information about internal/external links, internal/external objects, and other general site statistics seen at different levels of the website design. In total, the crawl

[3] Alexa, http://www.alexa.com, Apr. 2012.

[4] Selenium, http://seleniumhq.org/, March 2012



(a) Landscape overview

(b) Chinese cluster

(c) Specialized IDPs

(d) Facebook (popular IDP)

Fig. 2. Overview of the third-party identity management landscape. (Manual top-200 dataset from 2012.)

included the download of $1.4 \cdot 10^8$ objects (totaling 1.6 TB), the identification and analysis of $2.5 \cdot 10^7$ links. For all sampled sites and the sites with which the sampled sites have third-party relationships, we also used complementing scripts to obtain additional ownership and geographic information.

A. Landscape Overview

Figure 2 shows the RP-IDP relationships that we manually identified pertaining to the 200 most popular websites and all IDPs used by these sites. IDPs are red, RPs yellow, and hybrid websites (acting as both RP and IDP) are orange.

Note that the hierarchy is relatively flat with a few popular IDPs, such as Facebook (Figure 2(c)) dominating the usage, and very few hybrid websites (e.g., Yahoo). Yet, as illustrated by the Chinese cluster (Figure 2(b)), the existence of multiple hybrids can allow for more complex structures and increased depths. We have also observed that Chinese RPs almost exclusively select Chinese IDPs [11]. However, it is unclear how much of this is due to censorship-related restrictions.

The high usage of popular social IDPs such as Facebook, adds value to both the IDPs and the passwords associated with these IDPs. For example, even if the user does not have an account on a specific RP, an attacker knowing the IDP password of the user can easily impersonate the user on any RP using that IDP. IDPs with higher node degree may become increasingly targeted by potential attackers.

IV. Current State: RPs and their IDPs

A. Methodology Overview

We apply the following basic methodology. First, during data collection, we identify RP-IDP relationships and other site characteristics for selected sample websites. Second, we classify the sampled site along four dimensions (summarized and explained in the following section). Third, we use hypothesis testing to identify website classes more likely to act as RPs.

B. Website Classification

Primary service: We manually labeled the top-200 pages (and their IDPs) into one of eight service classes: *Social/portal*, *Tech*, *Commerce*, *News*, *Video*, *Info*, *Filesharing*, *Ads*, and *CDN*. The classes used here are more carefully explained in earlier work [2], [11].

Popularity segment: We classify websites based on which popularity segment they belong to after taking the logarithm of the Alexa site rank. For this study, we defined five classes: $[1, 100]$, $(100, 10^3]$, $(10^3, 10^4]$, $(10^4, 10^5]$, and $(10^5, 10^6]$. We apply logarithmic transformation because ranks follow power-law distributions [2], ensuring that each popularity segment is well represented.

Geographic region: While we have considered a wide range of geographic measures and our conclusions regarding potential geographic biases appear to hold true in general, for the purpose of this article we use the combination of two measures. First, we use a geo-location service to determine the location of the servers. Second, we use statistics provided by Alexa to estimate the region in which their primary user audience is located. While neither of these two measures provide an objective quantification of the number of sites that are from a particular region, both measures allow us to place each site in one of six regions: North America, Europe, Asia, South America, Africa, and Oceania. If both measures maps to the same region, we classify a page to *North America*, *Europe*, *Asia*, or *Other region*. Otherwise, we classify the page as *Mixed/unknown*.

Byte/link volume: We also classified sites based on quantifiable metrics such as the number of links, objects, and bytes associated with different depths of the website. We (i) performed a Principal Component Analysis (PCA)[5], (ii) identified two principal components (one "volume-based" and one "link-based") that are responsible for 90.8% of the variations, and (iii) determined that these components are dominated by the depth-two metrics. Motivated by these observations and inspired by Pareto's 80-20 rule, we (i) label the 20% sites with the most external links as L^+ (the rest of the sites as L^-), and (ii) label the 20% sites with the largest total volume of objects as V^+ (and the rest of the sites as V^-). This labeling results in four classes: (V^+, L^+), (V^+, L^-), (V^-, L^+), and (V^-, L^-).

C. Who are the RPs?

Table I(a) shows the site types that are the most frequent RPs in the manual and crawled datasets, respectively, as well as for the subset of the crawled dataset that corresponds to the top-200 sites. Table I(b) shows the site types with the relatively highest portion of RPs, for the same datasets.

The data for the top-200 sites help validate the crawled dataset. For example, 7 out of 8 of the top categories are the same for the top-200 sites (first two columns in each row). The results only differ for the geographic region with the largest fraction of sites being RPs (row three, Table I(b)). This difference is in part due to the crawler being somewhat more cautious in classifying some Asian websites. However, it should be noted that North America is only 3% behind Asia in the manual dataset. (In the crawl/top200 dataset Asia is 14% behind North America.) We have shown that the crawler carefully and successfully avoids false positives [11]. This is reflected in the lower number of RP sites (Table I(a)) and percentages of sites that are RPs (Table I(b)) observed using the crawler (column 2) relative to the manual dataset (column 1).

Comparing the results for the full crawled dataset and for the top-200 sites, we note that the RPs in the top-200 typically are lighter in their design, as indicated by the difference in byte/link volume. It should also be noted that the impact of popularity is best evaluated when using the full crawled dataset.

Motivated by the aforementioned observations, we focus the remaining analysis on the crawled dataset when available (popularity segment, geographic region, and byte/link volume), and the manual dataset otherwise (primary service).

We observe that the "typical" RP, as defined by the most frequently observed RP classes (Table I(a)), is often of the same classes as the "typical" (most frequently observed) sample site. For example, Social websites are the most common class on the top-200 list (84 out of 200), more websites maps to North America than any other region (8,188 out of 35,619), most websites belong to the long tail of less popular websites, and by definition most sites will be classified as (V^-, L^-). Perhaps more interestingly, the website classes that are most likely to use at least one third-party IDP (Table I(b)) differ from the most frequently observed RP classes.

Using our site classification, the classes most likely to be an RP are: News sites, sites among the top-1,000 most popular websites, sites located in Asia, and sites belonging to the (V^+, L^+) category of "heavy" sites with a lot of content (V^+) and links (L^+). Here, we only present our key findings and highlight the classes that showed significantly higher IDP usage, in particular the classes for which we have at least 99% confidence with our hypothesis tests.[6]

First, although the manual dataset, for which we manually label websites based on their primary service, only covers 200 pages, the much higher than average IDP usage among News sites is significant ($\geq$ 99% confidence). Out of the 17 News sites in the top-200 dataset, 11 (65%) acted as RP, which can be contrasted against the overall averages of 69/200 (35%) among the top-200 websites. In total, these News RPs had 29 RP-IDP relationships, with a total of 7 unique IDPs.

The News sites are interesting because they often use popular Social IDPs (e.g., Facebook and Twitter) to incentivize sharing content, comments, and popular articles, such as to increase readership. User information from these social IDPs may also be used to help personalize and improve the user experience; e.g., by providing targeted news stories and commercial ads. The last point is supported by the frequent use of an easy-to-integrate login-widgets from Gigya which allows easy collection of personal information from multiple IDPs.

Second, the popular sites (top-10,000 websites) have a significantly higher IDP usage. For example, using our large-scale (automatic) dataset, for which we conservatively identify that on average 5.2% of the sampled sites are RPs, we observe that among the top-1,000 websites 12.0% of the sampled sites are RPs. Also, the IDP usage in popularity segments $(10^3, 10^4]$ is significant, with 8.9% of the sites being identified as RPs. The significant skew towards popular sites using IDPs is interesting as a larger fraction of these sites appears up-to-date, using more Web 2.0 features, and may hence be earlier adopters. Third, both North American (6.7%) and Asian (8.4%) sites appear to be significant IDP users. Finally, the heaviest sites (with (V^+, L^+) or (V^+, L^-)) have significant (12.1% and 10.7%) IDP usage.

D. Who are the popular IDPs?

Table II summarizes the top-10 most popular IDPs in our dataset, the Alexa ranks of these services, the number of (sampled) RPs that these IDPs help, and the primary protocol used for third-party authentication. The low Alexa rank for vkontakte.ru is largely due to a domain name change (to vk.com, with an Alexa rank of 41), as many RPs use the old domain name.

[5] The PCA was based on six log-transformed metrics (the number of links, objects, and bytes associated with the first page, as well as the same metrics as measured down to a depth two from the first page), each with variances normalized to one.

[6] The details of the hypothesis testing can be found in our extended version: http://www.ida.liu.se/~nikca/papers/ic15details.pdf.

TABLE I
Overview of class-based IDP usage analysis.

	Class with most identified RP sites		
Criteria	**Manual**	**Crawl/top200**	**Crawled**
Primary service	Social (23)	Social (10)	-
Popularity segment	$(100, 200]$ (37)	$(100, 200]$ (14)	Unpopular
Geographic region	North America (30)	North America (18)	North America (546)
Byte/link volume	(V^-, L^-)(37)	(V^-, L^-)(13)	(V^-, L^-) (943)

(a) Overview of the website classes with most identified RP sites.

	Site class most likely to use at least one IDP		
Criteria	**Manual**	**Crawl/top200**	**Crawled**
Primary service	News (65%)	News (46%)	-
Popularity segment	$(100, 200]$ (37%)	$(100, 200]$ (14%)	$[1, 10^3]$ (12%)
Geographic region	Asia (39%)	North America (22%)	Asia (8.4%)
Byte/link volume	(V^-, L^-) (40%)	(V^-, L^-) (13%)	(V^+, L^+) (12%)

(b) Overview of the website classes most likely to use at least one IDP.

TABLE II
Top-10 list of global IDPs. ([a] Facebook is a well-known OAuth-only provider, but has in the past been an RP in OpenID. [b] Google and Yahoo also occasionally uses OAuth. [c] The OpenID field allows general login with any OpenID IDP, although some restrictions may occur.)

IDP rank	**Alexa rank**	**IDP/federation**	**Protocol**	**Number of RPs**
1	2	facebook.com	OAuth[a]	1293
2	10	twitter.com	OAuth	378
3	9	qq.com	OAuth	278
4	1	google.com	OpenID[b]	250
5	4	yahoo.com	OpenID[b]	141
6	16	sina.com.cn	OAuth	127
7	-	openID	OpenID[c]	87
8	4173	vkontakte.ru	OAuth	73
9	25	weibo.com	OAuth	64
10	12	linkedin.com	OAuth	63

Use of popular sites as IDPs is dominant because these already have a large number of users with active accounts. In addition, in many cases, these sites may already have access to large amounts of personal information that could help the RP improve their personalization and service. No specialized IDP makes the top-10 list, the OpenID usage is small, and the use of the general OpenID field, which allows the user to input any OpenID provider, is only used by 87 of the 35,619 sampled sites (and 1,865 RPs). Instead, OAuth is the dominating protocol. Among the top-10 IDPs, eight IDPs use OAuth as their primary protocol and nine use OAuth for some of their relationships. The high OAuth usage can lead to significant cross-cite information leakage, with privacy implications for the users [12].

E. The Evolving RP-IDP Landscape

There are a few dominant IDPs that are widely used across almost all classes of websites. For example, Facebook is the most popular IDP in 19 of the 21 single dimension website classes defined in Section IV-B. (Tech and Asian sites being the only exceptions.) In addition to acting as an IDP for many websites, these popular IDPs often have access to much personal information, making them a potential security and privacy concern for the users. As information is shared with ad services these concerns will continue to increase. However, the high popularity and visibility of these IDPs ensures that they are under the society's scrutiny, as potential security flaws and information leakages can damage their reputation.

There are also some significant differences in IDP selection. Most obvious are Tech sites which use specialized IDPs to a larger extent than others and is the only class with MyOpenID and Yahoo among the top-3 IDPs. Also the sites mapping to Asia stands out, having QQ as their primary choice. The Asian Web also include more hybrid websites, acting as both RP and IDP, than observed within any other category. While Facebook, Twitter, and Google dominates the top-3 positions for most classes (19/21, 15/21, and 14/21, respectively), QQ also shows up in the 9 out of 21 (9/21) top-3 lists.

Increasingly many RPs select popular IDPs solutions that offer OAuth-based services that allow the RP to perform actions on the IDP and share information across sites, on behalf of the user. This trend has further increased the skew towards high-degree nodes such as Facebook (e.g., Figure 2(c)). For example, between Apr. 2012 and Feb. 2014 the use of Facebook increased from 44 to 52 RPs (18%) on the top-200 list. Similarly, Twitter and QQ increased from 15 to 19 (27%) and from 9 to 14 (56%), respectively. The biggest upswing was observed for Sina/Weibo, who increased from 5 to 13 RPs (160%).

While we have observed a small cluster of professional IDP service (Figure 2(d)), used mostly by technical top-200 sites, this cluster have been shrinking with time. For example, over the duration of the study presented here, the number of specialized RP-IDP relationships among the top-200 sites have decreased from 13 to 3 between Apr. 2012 and Feb. 2014; 8 to 1, if we do not count the general OpenID field. In fact, during this time, ClaimID, Vidoop, Clickpass, and MyOpenID have all gone out of business. Verisign's PIP service (now owned by Symantec) is the only specialized pro-IDP that we have observed among the top-200 websites, which will remain at least for the near future.

Another interesting trend is that the Chinese cluster (Figure 2(b)) has (i) become even more clustered (with the addition of more relationships and hybrid nodes, and (ii) further disconnecting from the other top-200 sites. While the original dataset had 6 relationships to IDPs outside the cluster (all to Live, owned by Microsoft), 5 of these 6 relationships were dropped around the time of the Snowden scandal in late 2013.

Finally, the negligible use of federated protocols [6] and a high skew in IDP selection effectively results in a single point of failure, with both the users and RPs becoming dependent on the IDPs' authentication services being available 24-7. Also, with these IDPs often being used for fast-and-easy every-day login to social networks, the users may not realize the full risks associated with these passwords.

V. Concluding Remarks

This article presents a class-based analysis that characterizes the third-party IDP landscape.

At a high level, the "typical" (most frequently observed) RPs are North American sites and often belong to the long tail of less popular sites. In contrast, the sites that are the most likely to act as RPs are News sites, sites that are among the 1,000 most popular sites on the Web, and sites that are hosted in and are serving an Asian audience. Using hypothesis testing we have found that these sites show a significant bias for using IDPs, relative to other sites. We have also found that most classes of sites, when selecting to act as RP, have a similar IDP usage and are often dominated by heavy Facebook usage. Except Asian sites, which are QQ dominated, and Tech sites, which uses specialized IDPs to a larger extent than others, most other classes differ primarily in their secondary IDPs. For example, the common use of a Gigya widget among News sites results in a more balanced IDP usage than seen in other classes. These three last mentioned classes (Asian, Tech, and News sites) are particularly interesting as they appear to be among the early adopters, but all substantially differ in their characteristics and IDP selection behavior.

We observe heavy bias, across website classes, towards selecting popular social IDPs, such as Facebook, Twitter, and QQ. The heavy skew in usage towards a few popular IDPs places high trust on these IDPs. Not only does a compromised IDP account give access to the user's RP accounts, but the broad usage of these IDPs also indirectly allows an attacker to easily and quickly impersonate the user across all types of Web services. What makes these third-party services attractive for the RPs, may also make the set of RP-IDP relationships attractive to the attackers!

References

[1] R. Dhamija and L. Dusseault. The seven flaws of identity management: Usability and security challenges. *IEEE Security & Privacy*, 6(2):24 – 29, 2008.

[2] P. Gill, M. Arlitt, N. Carlsson, A. Mahanti, and C. Williamson. Characterizing organizational use of Web-based services: Methodology, challenges, observations, and insights. *ACM Transactions on the Web (TWEB)*, 5(4):19:1–19:23, 2011.

[3] D. P. Kormann and A. D. Rubin. Risks of the passport single signon protocol. *Computer Networks*, 33(1-6):51 – 58, 2000.

[4] S. Landau, H. Gong, and R. Wilton. Achieving privacy in a federated identity management system. In *Proc. FC*, 2009.

[5] S. Landau and T. Moore. Economic tussles in federated identity management. *First Monday*, 17(10), 2012.

[6] L. Lynch. Inside the identity management game. *IEEE Internet Computing*, 15(5):78 – 82, 2011.

[7] M. Shehab, S. Marouf, and C. Hudel. ROAuth: Recommendation based open authorization. In *Proc. SOUPS*, 2011.

[8] S.-T. Sun and K. Beznosov. The devil is in the (implementation) details: an empirical analysis of OAuth SSO systems. In *Proc. ACM CCS*, 2012.

[9] S.-T. Sun, Y. Boshmaf, K. Hawkey, and K. Beznosov. A billion keys, but few locks: The crisis of Web single sign-on. In *Proc. NSPW*, 2010.

[10] S.-T. Sun, K. Hawkey, and K. Beznosov. Systematically breaking and fixing openid security: Formal analysis, semi-automated empirical evaluation, and practical countermeasures. *Computers & Security*, 2012.

[11] A. Vapen, N. Carlsson, A. Mahanti, and N. Shahmehri. Third-party identity management usage on the Web. In *Proc. PAM*, 2014.

[12] A. Vapen, N. Carlsson, A. Mahanti, and N. Shahmehri. Information sharing and user privacy in the third-party identity management landscape. In *Proc. IFIP SEC*, 2015.

[13] R. Wang, S. Chen, and X. Wang. Signing me onto your accounts through Facebook and Google: a traffic-guided security study of commercially deployed single-sign-on Web services. In *Proc. IEEE Symposium on S&P*, 2012.