

Third-party Identity Management Usage on the Web

Anna Vapen¹, Niklas Carlsson¹, Anirban Mahanti², and Nahid Shahmehri¹

¹ Linköping University, Linköping, Sweden

² NICTA, Sydney, NSW, Australia

Abstract. Many websites utilize third-party identity management services to simplify access to their services. Given the privacy and security implications for end users, an important question is how websites select their third-party identity providers and how this impacts the characteristics of the emerging identity management landscape seen by the users. In this paper we first present a novel Selenium-based data collection methodology that identifies and captures the identity management relationships between sites and the intrinsic characteristics of the websites that form these relationships. Second, we present the first large-scale characterization of the third-party identity management landscape and the relationships that makes up this emerging landscape. As a reference point, we compare and contrast our observations with the somewhat more understood third-party content provider landscape. Interesting findings include a much higher skew towards websites selecting popular identity provider sites than is observed among content providers, with sites being more likely to form identity management relationships that have similar cultural, geographic, and general site focus. These findings are both positive and negative. For example, the high skew in usage places greater responsibility on fewer organizations that are responsible for the increased information leakage cost associated with highly aggregated personal information, but also reduces the user's control of the access to this information.

1 Introduction

With an increasing demand for personalized services, many websites ask their users to create personal user accounts and authenticate themselves before service. To simplify account creation and increase personalization opportunities, many sites use third-party identity management services. These services allow a user's digital identity and some personal information to be shared across multiple distinct sites; however, they also come with their own reliability, privacy, and security concerns [9, 13].

Third-party identity management providers typically offer simplified authentication, using a *single-sign-on* (SSO) [13] service. In a typical authentication scenario, the browser of a user wanting to use such identity service interacts with two additional parties: a *relying party* (RP) and an *identity provider* (IDP). An RP such as Yahoo can offer that users use their account with a third-party IDP such as Facebook or Google, to login to Yahoo and access its services. In this case, the selected IDP, say Facebook, would provide the authentication service and the user would only need to remember their digital identity with Facebook to access Yahoo. More formally, we say that two sites have an RP-IDP relationship, if the user log in to one of the sites (the RP) using the other site (the IDP). Figure 1 illustrates a more complex scenario. The Huffington

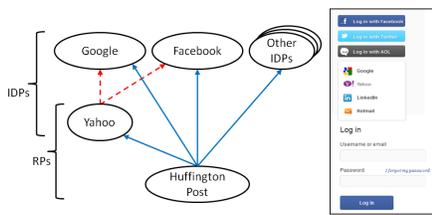


Fig. 1. Huffington Post login example.

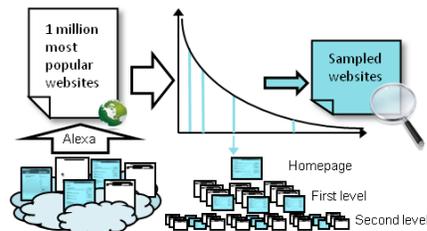


Fig. 2. Methodology overview.

Post site allows local login as well as authentication at multiple trusted third-party IDPs, including both Facebook and Google. Notice that Huffington Post uses Yahoo as IDP, illustrating that Yahoo can act as both IDP and RP. While the figure only shows a snippet of this complex situation, clearly, the relationships between these sites are nested.

In this paper we consider any third-party login collaborations in which an RP is using one or more external IDPs in the login process. In addition to SSO service, third-party identity providers are increasingly also used to share and modify information across sites. Following this trend, in addition to authentication protocols, such as OpenID³ (e.g., used in part by Google), our study shows that these services are increasingly implemented using authorization protocols, such as OAuth⁴ (e.g., used by Facebook). Authorization protocols have the added functionality that they, on behalf of the user, can allow one site (or service) to perform actions on a different site.

In this paper, we present (i) a novel Selenium-based data collection methodology that allows us to accurately identify and validate RP-IDP relationships that are not easily captured by pattern-matching crawlers (Section 2), and (ii) a large-scale characterization of the identified RP-IDP relationships⁵, studying how websites select their IDPs (Section 3). Our methodology and analysis capture the impact on IDP selection of things such as relative site popularity, cultural/geographic biases, and the intrinsic website characteristics. To put distribution and selection characteristics in perspective, we compare our results against what is observed for third-party content delivery relationships. To the best of our knowledge, at the time of writing, there are no other large-scale studies of the third-party identity management landscape and its structure.

In general, we find that IDP usage is highly skewed, with a small set of IDPs accounting for most of the world-wide usage. These IDPs are typically globally popular sites (such as Facebook, Twitter, and Google) that have a large user base. The tendency to select popular services as IDPs has resulted in a pronounced rich-gets-richer effect. For example, 90% of the RP-relationships are to an IDP among the top-100 most popular sites according to Alexa, and 50% of the observed RP-IDP relationships have a site-rank ratio of at least 10^3 (ratio between the RP and IDP ranks). In contrast, the third-party content provider relationships are relatively evenly distributed among Alexa ranks, with 50% of the observed content provider relationships having a site-rank ratio of at least four, suggesting a relatively weaker bias towards more popular sites. We also find that IDP selection is more biased towards the same cultural/geographic region than

³ OpenID (official website), <http://openid.net/>, May 2013.

⁴ OAuth (official website), <http://oauth.net/>, May 2013.

⁵ Datasets are available at <http://www.ida.liu.se/divisions/adit/data/pam14.html>.

what is observed for third-party content provider selection, and websites appear to have a slight preference for selecting IDPs that provide similar service as the website.

2 Methodology

2.1 Data Collection

At the core of the third-party identity management landscape are the relationships between relaying parties (RPs) and identity providers (IDPs). To identify relationships for a wide range of sites, we employed a novel two-step approach. In the first step, we use a logarithmic sampling technique to pick a sample set of sites with varying popularities. In the second step, we use a Selenium-based crawling tool that allows us to accurately identify and validate RP-IDP relationships that are not easily captured by pattern-matching crawlers. Figure 2 summarizes our methodology.

Popularity-based logarithmic sampling: The size of the Web precludes identification of all RP-IDP relationships. Instead, we collect a “sample” set of sites and study these for potential IDP usage. For our sampling we retrieved the Alexa⁶ list on April 17, 2012 of the top 1 million most popular websites worldwide. Based on Web popularity following power-law distributions [5], we then placed 80,000 points uniformly on a logarithmic range $[1, 10^6]$ (ensuring that the number of points in segment $[10^x, 10^{x+1}]$ is independent of x) and sampled the sites with a popularity rank closest to each point. After removal of duplicates we had a sample set with 35,620 sites.

Relationship identification: To identify RP-IDP relationships of each sampled site, we built a novel Selenium-based⁷ crawling tool. Our crawler is built as a cloud-based proxy, and acts like a human user, which may click on many different available GUI elements, react to pop-ups, and take a range of other GUI-driven actions. In comparison with pattern-matching tools, we have found that our Selenium-based tool does a very good job identifying relationships associated with more complex Web 2.0 websites. It captures relationships that are non-trivial and not easily identified even using manual methods. This is important as many websites today use clickable images which may not match the actual IDPs, and/or hide the IDPs within their design.

Our crawler is multi-threaded, explores each site down to a depth two from the starting page, and uses regular-expression-based pattern matching on all elements (including not only links but also clickable images, pop-ups, and similar) to find elements related to authentication. The tool initially prioritizes clicking on object elements that match pre-defined patterns and therefore are potential authentication related elements. When all matched elements on a page are clicked, the tool continues clicking all other clickable elements until a specified per-site timeout value of 25 minutes is reached.

2.2 Statistics and Complementary Datasets

For both the sampled sites and the identified IDPs, we used the Selenium-based crawler and complementing scripts to collect statistics and information about each site. The

⁶ Alexa (official website), <http://www.alexa.com>, April 2012.

⁷ Selenium, <http://seleniumhq.org/>, March 2012.

crawl 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. Out of the 35,620 sampled websites, 1,865 websites were classified as RPs. We also observed 50 IDPs and 3,329 unique RP-IDP relations. Finally, complementing scripts were used to obtain additional ownership, cultural, and geographic information regarding all observed (sampled and non-sampled) sites. These complementary datasets are primarily supportive and are discussed when used.

2.3 Validation

To ensure a fairly clean dataset, our Selenium-based data collection tool is designed to carefully identify true RP-IDP relationships and avoid false positives. The accuracy of our tool was validated using semi-manual relationship identification and classification. To obtain as exhaustive and accurate a list of RP-IDP relationships as possible we first built yet another crawling tool that identified candidate relationships much more loosely (and hence resulted in many false positives!) that we could later manually verify/reject by examining the identified objects that suggested such a linkage. We also carefully explored all sites on the top-200 list manually for relationships. By combining these two approaches we built a list of relationships involving the top-200 sites.

Out of a total of 69 RPs, 32 IDPs, and 186 relationships, the tool identified 23, 12, and 36, respectively. While the tool clearly does not find all relationships, it is very successful in avoiding false positives. The single false positive for IDPs (`vkontakte.ru`), is due to a name change from `vkontakte.ru` to `vk.com`. The three potentially false RPs (`wordpress.com`, `uol.com.br`, and `onet.pl`) are all blog hosts, portals and website hosts. Their users create sub-sites which may allow third-party authentication. The third site `onet.pl`, also shares content with Facebook. Finally, out of the 13 false relationships, roughly half are due to sub-domain matches and name changes (as discussed above), and the rest are due to misclassified content provider relationships.

While omitted, it should be noted that our conclusions have been tested and validated using multiple crawlers. The use of our semi-manual dataset further strengthens our belief in the generality of our results. We have not found any major biases in the set of relationships included.

3 Characterization Results

3.1 The Big Players

Table 1 summarizes the top-10 globally most popular IDPs in our dataset, the number of (sampled) RPs that these IDPs help, and the primary protocol used for third-party authentication. For reference, we also provide the Alexa ranks of the services.

We note that some of the most popular sites on the Web also are the most used IDPs. 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. Interestingly, the general OpenID field that allows the user to input any OpenID provider is only used by 87 of the sampled sites and no specialized IDP makes the list. In fact, we observe that OAuth is the dominating protocol. Among the top-10 IDPs in Table 1, eight IDPs use OAuth as their primary protocol and nine use OAuth for some of their relationships.

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

Table 1. 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.)

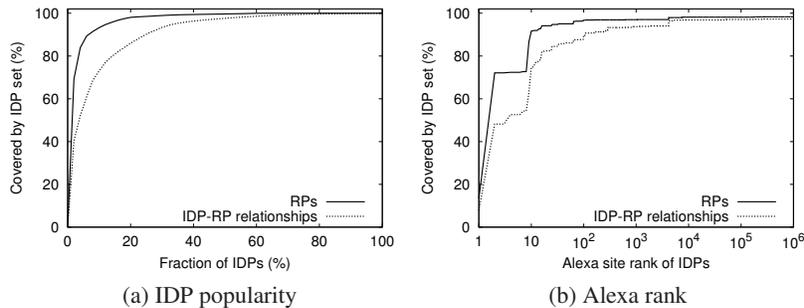


Fig. 3. RPs that are served by the most popular IDPs.

These results suggest that many sites choose to use popular sites as their IDPs. The users are more likely to already have accounts with these sites, and 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. This observation may also provide some insight as to why identity management federations such as OpenID and third-party services that specialize only in identity management have struggled to take off [11].

We next take a closer look at the relative popularity of the IDPs. Figures 3(a) and 3(b) show the fraction of RPs that are served by the most popular IDPs and the IDPs of a certain global popularity, respectively. We note that more than 75% of the RPs are served by 5% of the IDPs, and the majority of these 75% are made up by the IDPs with Alexa ranks in the top 100. In fact, only 15 of the 44 IDPs outside the top-10 Alexa list serve more than 10 sampled IDPs.

3.2 IDP Usage

We next consider the IDP usage. Figure 4 shows the number of IDPs observed for each sampled site. While the average is highest for the most popular sites, we note that there are some less popular sites that use a large number of IDPs. Among the nine sampled

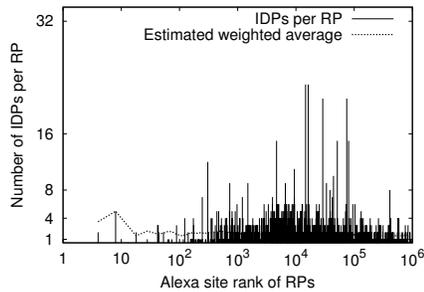


Fig. 4. Number of IDPs per sampled RP.

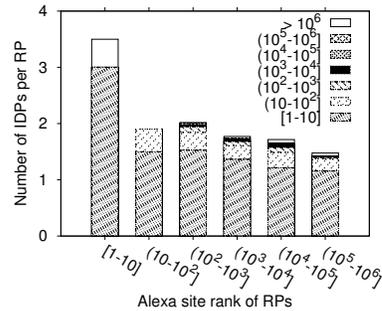


Fig. 5. Breakdown of the average number of IDPs selected per RP and popularity segment.

websites with more than ten IDPs, six of the sites are news sites and all nine use a login widget from `gigya.com`, providing the sites with a selection of IDPs.

Figure 5 breaks down the IDP usage for each popularity segment, based on which IDPs the sites in each segment are using. We can see that the RPs with the most popular sites on average use the most IDPs, and that the top-ranked IDPs are the most popular IDP choices for sites belonging to all popularity segments. For all segments, the IDPs that rank in the top-10 contribute for more than 75% of the IDP usage.

3.3 Comparison with Content Services

To put popularity skew and biases in perspective, in the following we compare our observations with those observed in the more traditional content delivery context. First, we consider the site-rank of the biggest *service providers* and *service users*. In the context of identity management, these entities correspond to the third-party IDPs and the RPs, respectively. In the context of content delivery, these entities correspond to the third-party content providers that deliver the content and the site that the content is delivered on behalf of, respectively. This comparison provides a natural reference point, as both IDPs and third-party content providers serve clients on behalf of the origin site.

Figure 6(a) shows the service provider breakdown; i.e., the fraction of user sites (RPs, for example) that are served by each third-party service provider (IDP, for example) of varying global Alexa rank. Figure 6(b) shows the service user breakdown; i.e., the fraction of third-party relationships that these sites are responsible for. We note that the content provider usage is distributed much more evenly across popularities than the IDP usage, which is heavily skewed towards the most popular sites.

In general, we find that IDPs often provide service for less popular RPs, whereas in the context of content delivery, it is much more common that the third-party content is served by less popular sites. While these less popular sites in some cases are backed up by a big company, the differences are striking. Figure 7 shows the relative difference in site rank between providers (e.g., IDPs) and service user sites (e.g., RPs) for the two types of relationships. These results show that the identity management landscape is significantly more skewed towards the big players than the content delivery ecosystem. For example, while 50% of the observed content provider relationships have a site-rank ratio of at least 4 (suggesting only a light bias towards more popular sites), the corresponding RP-IDP site-rank ratio is at least 10^3 (between the RP and IDP ranks).

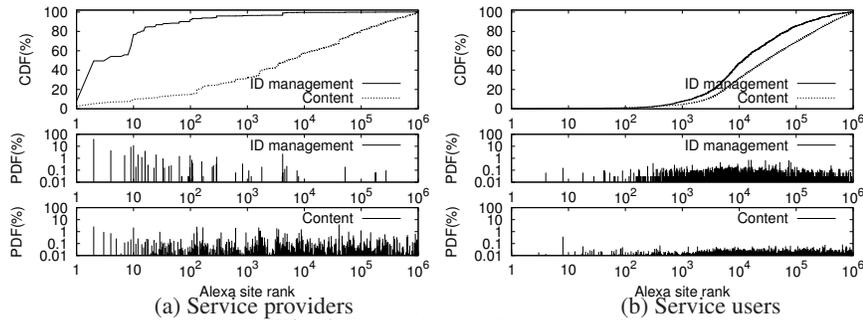


Fig. 6. Comparison with content delivery.

Type	Sites		Relationships			
	Total	RPs	Total	Per page	Per RP	Breakdown
Social/portal	84	23 (27%)	55	0.65	2.39	47 social, 4 tech, 3 commerce, 1 info
Tech	24	8 (33%)	40	1.67	5.00	26 social, 12 tech, 1 commerce, 1 info
Commerce	20	5 (25%)	6	0.30	1.20	3 social, 2 commerce
News	17	11 (65%)	28	1.65	2.55	28 social
Video	18	8 (44%)	24	1.33	3.00	22 social, 1 commerce, 1 info
Info	14	7 (50%)	11	0.79	1.57	10 social, 1 commerce
Filesharing	12	7 (58%)	22	1.83	3.14	22 social
Ads	6	0 (0%)	0	0	0	-
CDN	5	0 (0%)	0	0	0	-

Table 2. Manual site classification results for top-200 list.

3.4 Service-based Analysis

To gain a better understanding of the sites that are more likely to act as an RP or IDP, we manually classified the top-200 sites, as well as each of the identified IDPs, based on the primary service they provide. For this analysis, we manually labeled each site into one of nine service classes. While alternative classifications are possible, the classes used here were inspired by those used by Gill et al. [5]. Table 2 lists the service classes and the statistics for each class.

These results show that the use of IDPs is greatest among sites that share information/news/data. This is consistent with significant use of OAuth. However, it is also interesting to see that sites of some other service classes (e.g., tech and video sites) often use more IDPs per RP, in the case that they decide to act as an RP. We conjecture that the high number of IDPs per RP for tech sites reflects that these sites are early adopters of these technologies. This conjecture is supported by the fact that these sites to a much larger extent than other sites give their users the option of using specialized IDPs, which provide identity management as their only service.

We find that both tech and commerce sites have a relative preference for picking IDPs from within their own category. For example, the tech sites are responsible for 12/40 (30%) of the tech IDP relationships observed from the top-200 sites, and 2/5 (40%) of the IDPs used by RPs classified as commerce sites are to IDPs that are commerce sites. These sites may have to rely more on domain knowledge to maintain credibility within their communities than other type of sites. News and file sharing sites, on the other hand, only use IDPs classified as social (e.g., Facebook and Twitter). This

Region	Method (local (%))					
	Whois		Servers		Audience	
	IDPs	CPs	IDPs	CPs	IDPs	CPs
North America	97.0	91.7	95.8	88.8	95.8	82.2
Europe	0.4	21.2	0.9	21.5	0.9	22.0
Asia	61.3	25.7	71.8	45.5	53.6	41.3
Others	0.0	9.1	0.0	10.6	1.3	9.7

Table 3. Percent (%) unique third-party relationships that are to a local IDP or content provider (CP) in the same geographic region as the sampled site, using each of our three location mappings.

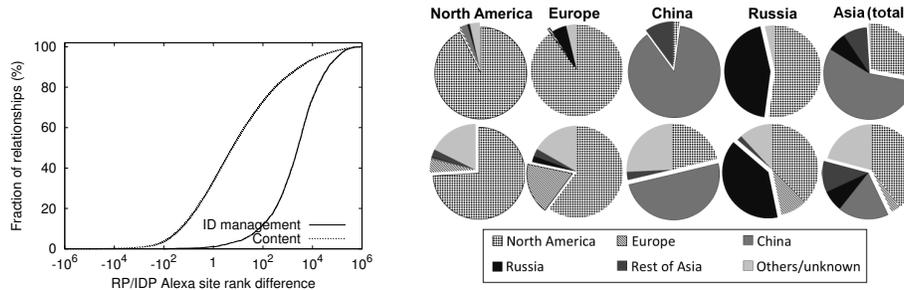


Fig. 7. Difference in site-rank ratio between service user/provider. (Alexa rank of user divided by rank of provider.)

Fig. 8. Geographic distribution of third-party relationships. Top row: Identity management. Bottom row: Content delivery.

usage may be motivated by a desire to form “personal” relationships and connect with more users.

3.5 Cultural and Geographic Analysis

We next try to glean some insight as to whether there may be some preference for selecting IDPs with similar geographic or cultural focus. Due to factors such as global user populations, it is difficult to uniquely assign each site to a single geographic region. We make no claims regarding the absolute number of sites that pick a “local” IDP. Instead, our observations are discussed relative to what is observed for third-party content delivery relationships.

While the general conclusions of our results appear to hold true for a wide range of mapping approaches, in this section we present results using three diverse methods: (i) a whois services based on where sites are registered, (ii) an online geo-location service located in the US to map the location of the servers, and (iii) statistics provided by Alexa estimating the region in which the site’s primary user audience is located.

Table 3 shows the percentage of local relationships, for each of the three mapping approaches. We use one row for sample sites mapped to different parts of the world, and list how large a percentage of the third-party providers are considered local. The region “others”, which includes South America, Africa and Oceania, is only responsible for 2-3.5% of the relationships. In the above results we exclude the unmapped relationships.

While the correct geographic location of a service/site in non-trivial and the exact percentage of sites classified as “local” clearly depends on the method used, our results allow two major observations. First, we note that there is a tendency for selecting local

IDPs in all geographic regions with major IDPs, with the exception of Europe and Other, which primarily use major American IDPs. Second, and more importantly, the fraction of “local” IDPs is larger than the fraction of “local” content providers for all regions except for Europe and Other. It should be noted that the IDP usage in Europe and Other overall is much smaller than in the other regions. This stronger locality preference (with the exception of European sites) is further illustrated in Figure 8, which shows the geographic breakdown for sites mapped to the regions with the most usage. Here, the server-based mapping approach is used and we include a further breakdown of Asia.

A closer look at the data reveals that info RPs only choose local IDPs, whereas social, tech, file sharing, and news RPs use more non-local IDPs than other categories.

4 Related Work

Other works have formally validated OAuth [3,8], Facebook Connect [7], OpenID [12], and SSO services in general [1]. For OAuth, a long laundry list of problems and threats has been identified, including phishing, eavesdropping, and various problems related to tokens [6]. It has also been shown that SSO services are vulnerable to cross-site scripting attacks [1] and that OpenID is vulnerable to cross-site request forgery attacks [12]. However, perhaps the biggest weakness in SSO is the implementation of the protocols themselves [3, 10]. For example, Wang et al. [13] presents an experimental study of the security of some of the most popular SSO services.

It should also be noted that identity management solutions can greatly affect user behavior. For example, as users get used to authenticating with unknown third-parties and following the path of least resistance, they may not take security precautions or read privacy agreements, making them increasingly susceptible to phishing attacks [4]. This last uncertainty has prompted many users to be cautious and afraid of using (unknown) third-party services, and may be another reason why we observe that many sites select popular IDPs. This shift away from specialized OpenID solutions may also hamper the development and/or adoption of large-scale identity federations with stronger authentication methods [2].

Complementing prior work, we provide a large-scale characterization in which we analyze relationships and third-party selection in the identity management landscape.

5 Discussion and Conclusions

This paper studies how websites are using third-party identity providers. We present a novel data collection methodology, which combines a Selenium-based crawler and a log-based sampling technique, and use the collected datasets to characterize global IDP usage. To provide a reference point for discussion, we compare our observations with the selection of third-party content providers. Our methodology and analysis captures how factors such as relative site popularities, cultural/geographic biases, and the intrinsic characteristics of the websites influence the relationships between RPs and IDPs.

Our study shows that a small number of IDPs dominate the IDP space. These IDPs are typically popular web services with a large international user base, and already have access to large amounts of sensitive user data. As these companies are under public

scrutiny, they will hopefully take greater care to securely store and handle sensitive user information. However, the many RPs using a few IDPs can also result in a large attack surface. The fact that OAuth is dominating OpenID also raises some privacy concerns, as it typically involves sharing of more user data.

A rich-gets-richer phenomena appears to be at play with the tendency of RPs selecting highly popular services as IDPs. Instead of picking specialized IDPs, which provide authentication as their primary service, RPs are choosing IDPs with higher popularity ranking than themselves. Today, many specialized IDPs, such as Clickpass, Vidoop, and MyopenID are therefore disappearing, being acquired, and/or going out of business.

Our characterization and dataset is the first large-scale measurement-based study of the identity management landscape and its structure, and is expected to provide an important stepping stone towards better understanding third-party identity management and their impact on Web users. Future work includes a large-scale security and privacy evaluation of alternative identity management solutions that take into account the observed relationship tendencies observed in the current identity management landscape.

References

1. A. Armando, R. Carbone, L. Compagna, J. Cuellar, G. Jorge, G. Pellegrino, and A. Sorniotti. From multiple credentials to browser-based single sign-on: Are we more secure? In *Proc. IFIP SEC*, June 2011.
2. J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Proc. IEEE Symposium on S&P*, May 2012.
3. S. Chari, C. Jutla, and A. Roy. Universally composable security analysis of oauth v2.0. Technical report, Cryptology ePrint Archive, Report 2011/526, 2011.
4. R. Dhamija and L. Dusseault. The seven flaws of identity management: Usability and security challenges. *IEEE Security & Privacy*, 6(2):24–29, Mar/Apr. 2008.
5. 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, Oct. 2011.
6. T. Lodderstedt, M. McGloin, and P. Hunt. OAuth 2.0 threat model and security considerations. Internet-Draft, IETF, Oct. 2011.
7. M. Miculan and C. Urban. Formal analysis of facebook connect single sign-on authentication protocol. In *Proc. SOFSEM*, 2011.
8. S. Pai, Y. Sharma, S. Kumar, R. M. Pai, and S. Singh. Formal verification of oauth 2.0 using alloy framework. In *Proc. CSNT*, June 2011.
9. B. Pfitzmann and M. Waidner. Analysis of liberty single-sign-on with enabled clients. *IEEE Internet Computing*, 7(6):38–44, Nov. 2003.
10. S.-T. Sun and K. Beznosov. The devil is in the (implementation) details: an empirical analysis of oauth sso systems. In *Proc. ACM CCS*, Oct. 2012.
11. 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*, Sept. 2010.
12. 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.
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*, May. 2012.