Information Sharing and User Privacy in the Third-Party Identity Management Landscape

Anna Vapen, Niklas Carlsson, Anirban Mahanti, Nahid Shahmehri

Information Sharing and User Privacy in the Third-party Identity Management Landscape

Anna Vapen[1], Niklas Carlsson[1], Anirban Mahanti[2], and Nahid Shahmehri[1]

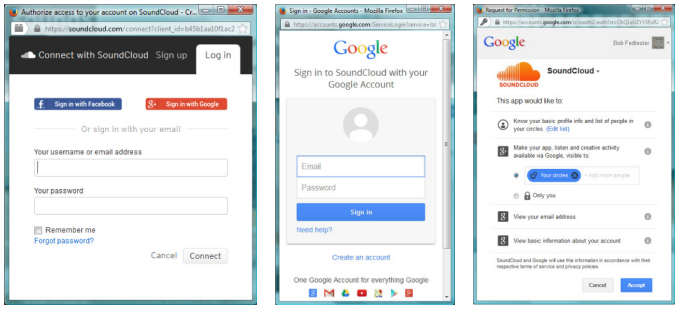
[1] Linköping University, Sweden, firstname.lastname@liu.se
[2] NICTA, Australia, anirban.mahanti@nicta.com.au

Abstract. The cross-site information sharing and authorized actions of third-party identity management can have significant privacy implications for the users. In this paper, we use a combination of manual analysis of identified third-party identity management relationships and targeted case studies to (i) capture how the protocol usage and third-party selection is changing, (ii) profile what information is requested to be shared (and actions to be performed) between websites, and (iii) identify privacy issues and practical problems that occur when using multiple accounts (associated with these services). By characterizing and quantifying the third-party relationships based on their cross-site information sharing, the study highlights differences in the privacy leakage risks associated with different classes of websites, and provides concrete evidence for how the privacy risks are increasing. For example, many news and file/video-sharing sites ask users to authorize the site to post information to the third-party website. We also observe a general increase in the breadth of information that is shared across websites, and find that due to usage of multiple third-party websites, in many cases, the user can lose (at least) partial control over which identities they can merge/relate and the information that is shared/posted on their behalf.

1 Introduction

Many popular web services, such as Facebook, Twitter, and Google, rely heavily on their large number of active users and the rich data and personal information these users create or provide. In addition to monetizing the high service usage and personal information, the rich user data can also be used to provide personalized and customized user experiences that add value to their services.

With this in mind, it is perhaps not surprising that many websites are using third-party single sign-on (SSO) [5, 14] services provided by popular websites. With SSO, a website such as Soundcloud will partner with one or more other third-party websites (e.g., Facebook and Google), which will be responsible for user authentication on behalf of Soundcloud. As illustrated in Figure 1(a), a user is given the option of using Facebook and Google for authentication. Assuming that the user selects Google, the user is redirected to Google for authentication (Figure 1(b)). We refer to Soundcloud as a relying party (RP) and Facebook/Google as a third-party identity provider (IDP).



(a) IDP selection (b) Authentication (c) App rights

Fig. 1. Soundcloud example illustrating the login process when using IDPs, as well as the app rights requested by Soundcloud when using Google as IDP.

In addition to providing an authentication service, at the time of account creation or first login, the user is typically asked to approve an app-rights agreement (e.g., Figure 1(c)) between the user and the RP, which (i) gives permission to the RP to download agreed-upon information from the user's IDP account, and (ii) authorizes the RP to perform certain actions on the IDP, such as posting information. Such permissions place great responsibility on the RPs, and can raise significant privacy concerns for users.

Privacy concerns related to the rich information shared across websites will likely increase as more sophisticated statistical methods are used to reveal private information using only public data [10, 18, 3]. For example, public Twitter feeds can be used to determine political views and ethnicity [10], users can be identified across websites even when they lie about their identity [18], and even relatively innocent information such as the music that people listen to can reveal personal information many users want to keep private [3]. Despite many interesting case studies of how this information can be misused, and our previous basic characterization of the geographic and service-based biases in how RPs select their IDPs [15], we are not aware of any study that characterizes the cross-site information sharing and RP authorization in this third-party landscape. Such a study is important to assess the current privacy risks.

This paper provides the first broad empirical analysis of the cross-site information sharing, app-rights agreements, and account management complexities in the current third-party identity management landscape. We place particular attention on the personal information shared and accessed between different sites, and discuss potential privacy implications for end users. Motivated by a high skew in website popularities [7], our analysis primarily focuses on the cross-site information sharing associated with the 200 most popular websites.[3] Focusing on these sites allows us to manually identify and analyze the information sharing, account creation process, and website interactions as observed by the user.

[3] Alexa (official website), `www.alexa.com`.

1.1 Contributions and Roadmap

A high-level summary of the results were reported in a short (3-page) poster paper [16]. Here, we first briefly present a high-level characterization of the protocol and IDP usage observed in the wild (Section 2). Our results confirm that the use of authorization protocols such as OAuth is significantly more common than the use of pure authentication protocols such as OpenID and that OAuth is becoming increasingly dominant. Since OAuth allows richer cross-site information sharing and can authorize RPs to perform actions on the IDPs, on behalf of the user, these results reaffirm the importance of characterizing today's app-rights agreement usage and their privacy risks. We find that many RPs likely select IDPs on other criteria than protocol compatibility. We also find that some IDPs are much more frequently used in combination than on their own, and that there is a high concentration in IDP usage. This can be a potential privacy and security concern in itself, as the user credentials for the most popular IDPs become more valuable to an impersonator, for each RP that adds that IDP.

Thereafter, we characterize the cross-site information sharing and authorized app rights associated with the most popular IDPs (Section 3), responsible for the majority of RP-IDP relationships. We categorize app-rights agreements based on the type of information and actions granted to the RPs (Section 3.1) and identify the most commonly observed privacy risk categories (Section 3.2). We find significant differences in the information leakage risks associated with different RP classes (Section 3.3) and IDPs (Section 3.4).

Finally, we use targeted login and account creation tests to analyze the information sharing in scenarios in which the users have accounts with multiple IDPs (Section 4). Among other things, we find that there often is significant overlap in the information that an RP may request from different IDPs (Section 4.1), that there are significant differences in how well RPs combine the information from multiple IDPs (Section 4.2), and that some IDPs (e.g., Facebook) are much more likely to implicitly enable information leakage between IDP accounts through the RP than others (Section 4.3). In many cases, we have also found that the results depend on the order IDPs are added and that users often lose control of what is shown in their public profile with the RP. We even found several cases in which account merging is not possible or additional IDPs cannot be added/used.

The significant differences observed from case to case also illustrate that there is no common API, and that RPs typically pick IDPs based on IDP popularity and the information sharing and actions they enable, rather than protocol compatibility. With each RP implementing its own solution, the user must trust the RP and its implementation. Both OpenID and OAuth have many security issues [5, 1], even when used alone, especially if implemented incorrectly [17, 12]. Many privacy issues discussed here will therefore likely take time to address.

2 Protocol and IDP Selection

Today's RP-IDP relationships are typically implemented using OpenID or OAuth. While OpenID was designed purely for authentication and OAuth primarily is

an authorization protocol, both protocols provide an SSO service that allows the user to access the RP by authenticating with the IDP without needing local RP credentials. With OAuth, a local RP account is always linked with the users IDP account, allowing information sharing between the RP and IDP. Local RP accounts are optional with OpenID.

We primarily focus on all RP-IDP relationships that we have manually identified on the 200 most popular websites on the Web (as observed in Apr. 2012, Feb. 2014, and Sept 2014), but will also leverage the 3,203 unique RP-IDP relationships (3,329 before removing false positives) identified using our custom designed Selenium-based crawling tool [15].

OAuth is the dominant protocol as observed in both manual and crawled datasets. For example, in Apr. 2012, 121 of 180 (67%) relationships in the manual dataset and 2,543 of 3,203 (79%) relationships in the crawled dataset are directly classified as OAuth, compared to only 20 (11%) and 180 (6%) as OpenID relationships in the two datasets. Of the remaining relationships, 39 and 441 used an IDP that supports both OpenID and OAuth. Since then, as measured in Sept. 2014, we have seen a further increase of OAuth usage (+24%) and drop in OpenID usage (-10%) among the top-200 websites.

We have found that IDP selection differs significantly depending on how many IDPs an RP selects, and some IDPs are more likely to be selected together with others. In total the top-5 ranked IDPs are responsible for 92% (33 of 36) and 90% (1,111 of 1,233) of the relationships of RPs selecting one single IDP. For RPs with 2-3 IDPs, 83% and 75% of the relationships are to the top-5, but for RPs with 4 or more IDPs only 38% and 55% are to the top-5 IDPs. Facebook+Twitter is the most popular pairing with 37% (125 of 335) of all IDP pairs, Chinese QQ+Sina place second (19%), and Facebook+Google third (12%).

3 App Rights and Information Flows

This section considers the information sharing and actions that RPs are permitted. Although it is impossible to know exactly *how* each RP uses the information they obtain from the IDPs (e.g., if they use data mining to present targeted ads, provide better services, or if they simply store the information for potential future use), the app-rights agreements between RPs and users reveal (i) the information that the RP *will obtain* from the IDP, and (ii) the actions the RP *will be allowed* to perform on behalf of the user.

For this study, we have carefully recorded the app-rights agreements for the RP-IDP relationships identified in the manual top-200 dataset. We created fake accounts on the IDPs, initiated the account creation process for each identified RP-IDP relationship involving this IDP, and recorded the app-rights agreements that our fake users were requested to agree to. For the tests in this section, we interrupt the login process after recording the app rights. Due to limited translation support, we only recorded statistics that provided app-rights agreements in English. The use of fake identities helps remove potential biases in the app-rights agreements presented to the users.

A few IDPs required the use of phone numbers in the registration process. Although these phone numbers or advanced data mining techniques [18] can be used to link the fake profiles with real identities, we have not observed anything that would suggest that these phone numbers have impacted our results.

For this analysis we focus on the RP-IDP relationships in the top-200 dataset from Feb. 2014. However, as Facebook has significantly changed their API since then, from version 1.0 (Apr. 2010) to 2.0 (Apr. 2014), and again to 2.1 (Aug. 2014), we analyze recorded app rights from both Feb. 2014 and Sept. 2014.

3.1 Classification of Information

When analyzing app-rights restrictions as described in the APIs of the three major IDPs (Facebook, Twitter and Google) as well as the actual app-rights usage across the top-200 websites in our datasets, we have identified five different classes of app rights, each with their own privacy implications. The first four classes (B, P, C, and F) capture information (or content) transferred from the IDP to the RP. The last class (A) has to do with actions being performed by the RP, on the IDP, on behalf of the user.

- **Basic information (B):** Relatively non-private information that the user is often asked to provide websites, including identifiers (e.g., email address) to identify existing accounts, age range, language, and public information.
- **Personal information (P):** This class includes personal information, common in many basic "bundles" (e.g., gender, country, time zone, and friend list), but also more sensitive information such as political views, religion, and sexual orientation.[4]
- **Created content (C):** This class contains content directly or indirectly created by the user, and includes images, likes, and check-in history. The sensitivity of the data varies. For example, in some cases the user may want to share images and video across sites, while in other cases this content may be considered private. Also the sensitivity of "logged" content (e.g., likes, watched video clips, location history) varies significantly on the situation.
- **Friends' data (F):** This class consists of data of other users (e.g., friends of the user). Even when of a non-sensitive nature, this data is privacy critical since it belongs to another, potentially non-consenting, user.
- **Actions taken on behalf of the user (A):** This final class includes the right to actively export data from the RP to the IDP and the right to perform *actions* on behalf of the user. This may include posting information on a user's IDP timeline or feed. The transferred data may include content the user creates at the RP (e.g., images), or information about the user's actions on the RP (e.g., sharing read articles, or music the user has listened to).

[4] One current Facebook bundle and two Google bundles named "Basic information" and similar (from which the RP selects which bundle to use in the app-rights agreement presented to the user) include both class B and P information.

While the current example considers a scenario with a single RP-IDP pair, Section 4 briefly considers the multi-IDP case, in which information may be shared/leaked across multiple sites. Here, the action (A) class is particularly interesting when used in combination with friends' data (F). In this case, the RP may enable the actions by one user at one IDP to be leaked (through the RP) to other users at another IDP.

It is perhaps for this reason that Facebook, in their recent (Sept. 2014) multi-step app-rights agreements, does not share friend (F) data, and first request that the user approve data sharing permissions (B, P, and C), before the RP can ask the user to agree to optional action (A) permissions of different types. Regardless how these action permissions were classified in Feb. 2014, in the Feb. 2014 dataset, we include the most recent optional Facebook action (A) permissions from Sept. 2014. For the other big IDPs there have been no major changes in their APIs since Feb. 2014.

3.2 Risk Types

Today, many IDPs bundle the information requested into larger "bundles", and RPs must select which bundle to present to the users.[5] This simplifies the agreements, but reduces the granularity of control over information sharing, often resulting in the user being asked to grant permissions to share more information than the RP requires to perform the desired service.

Figure 2 summarizes all the observed app-rights agreements in our Feb. 2014 dataset. We use a Venn diagram to show all relationships involving actions in the left square and all others in the right square. Any relationship that is not in any of the three classes P, C, and F is in class B.

Only a handful of cases (4) limit themselves to only the basic (B) information (without actions (A)), and most RPs are requesting significantly more personal information from their users. These observations suggest that there is an expectation of trust in the RPs, beyond what the user would share publically. Generally, RPs that are performing actions (A) on behalf of their users are more likely to request access to content (C) from the IDP. In total, 40 of the 87 classified relationships include actions (A). Of these, 14 RPs also request access to content (C). Of the 47 app-rights agreements that does not request actions to be performed, only 12 (9+3) also request access to content (C).

Another important observation is that within each of the two boxes there is a clear ordering in risk types observed, suggesting that there is a natural ordering of the risk types observed in practice. In particular, class F is only used in combination with both C and P. This combination clearly has the highest privacy risks associated with it. Similarly, class C is only used in combination with P, clearly distinguishing its risks with those of sites that only request personal (P) or basic (B) information.

[5] For Twitter, the RP selected bundles (either "read", "read + write", or "read + write + direct messages") are translated into an explicit list of app rights presented to the users.

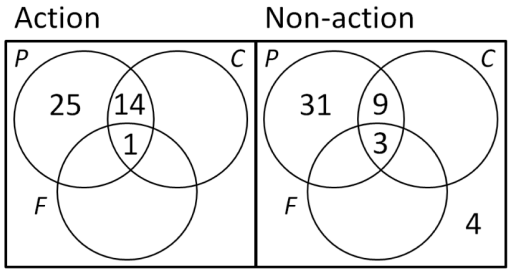


Fig. 2. RP-IDP relationships of different app-rights classes in the top-200 dataset.

Table 1. Risk types identified in dataset.

Risk type	Class combination	Risk type	Class combination
$\mathcal{R}_A^-$	$A \cap B$	$\mathcal{R}^-$	$\neg A \cap B$
$\mathcal{R}_A$	$A \cap P$	$\mathcal{R}$	$\neg A \cap P$
$\mathcal{R}_A^+$	$A \cap P \cap C$	$\mathcal{R}^+$	$\neg A \cap P \cap C$
$\mathcal{R}_A^{++}$	$A \cap P \cap C \cap F$	$\mathcal{R}^{++}$	$\neg A \cap P \cap C \cap F$

Motivated by these observations, we identify eight semi-ordered risk types (strict ordering within columns). Table 1 summarizes the observed risk types. We note that there is a strict privacy ordering in each column (from (-) to (++)), and with regards to each row (as allowing actions implies some risk), but that further ordering is not possible without making assumptions.

3.3 RP-based Analysis

Using the above RP-IDP relationship type classification, we next compare the app rights for different classes of RPs. In particular, we compare the app rights of RPs with different (i) primary web services, or (ii) number of IDPs.

Table 2 shows the number of relationships of each type, for websites that provides different web services. Here, we use a basic service categorization inspired by categories defined by Gill et al. [7]. With this categorization, each of the top-200 websites was manually classified as one of nine classes.

Among the classes with at least 10 RPs, News sites and File sharing sites are the most frequent users of actions (risk types $\mathcal{R}_A$ and $\mathcal{R}_A^+$), with 55% and 50% of their relationships including actions, respectively. Also Video sharing (67%) and Tech (63%) sites have a large fraction of relationships that include action (A) permissions. The high action (A) permission usage is likely an effect of these sites often wanting to promote articles, files, or videos to friends of the user. While we express privacy concerns regarding $\mathcal{R}_A^+$ relationships, these sites would in fact desire that the information that their articles/content are being read to propagate across many sites. This is also reflected in the relatively large number of IDPs per RP for these four website categories (3.3, 2.83, 2.33, and 3.0, respectively, compared to an overall average of 2.5).

Table 2. Breakdown of risk types of the RP-IDP relationships for RPs belonging to different websites categories, as classified based on their primary service.

	Sites	**Relationship type**							
Categ.	**RPs/Tot**	**Tested/Tot**	$\mathcal{R}^-$	$\mathcal{R}$	$\mathcal{R}^+$	$\mathcal{R}^{++}$	$\mathcal{R}_A$	$\mathcal{R}_A^+$	$\mathcal{R}_A^{++}$
Ads/CDN	0/9	-/-	-	-	-	-	-	-	-
Commerce	8/26	7/16	0	5	0	0	2	0	0
File sharing	6/10	12/17	2	3	1	0	3	3	0
Info	9/14	10/16	0	5	0	1	4	0	0
News	12/20	22/40	0	4	6	0	7	5	0
Social/portal	26/81	22/65	1	10	2	2	3	4	0
Tech	7/23	8/21	1	2	0	0	2	2	1
Video	9/17	6/21	0	2	0	0	4	0	0
Total	77/200	87/196	4	31	9	3	25	14	1

Table 3. Breakdown of risk types of the RP-IDP relationships for RPs with different numbers of IDPs.

		Relationship type							
IDPs	**RPs**	**Tested/Tot**	$\mathcal{R}^-$	$\mathcal{R}$	$\mathcal{R}^+$	$\mathcal{R}^{++}$	$\mathcal{R}_A$	$\mathcal{R}_A^+$	$\mathcal{R}_A^{++}$
1	36	24/36	0	11	3	2	7	1	0
2	15	19/30	1	7	0	1	7	2	1
3	11	21/33	1	6	3	0	6	5	0
4+	15	23/97	2	7	3	0	5	6	0
Total	77	87/196	4	31	9	3	25	14	1

We next compare RPs with different numbers of IDPs (Table 3). Interestingly, relationships including actions ($\mathcal{R}_A$ and $\mathcal{R}_A^+$) are primarily associated with RPs that have many IDPs. For example, while RPs with a single IDP use actions in 33% (8 of 24) of their relationships (all using Facebook), RPs with multiple IDPs (2, 3, or 4+) use actions in 48-53% of their relationships. As with our discussion about News sites and File sharing sites, the many IDPs of these RPs increases the risk for cross-site information leakage.

The most restrictive type ($\mathcal{R}^-$) includes only OpenID relations. For content sharing without actions ($\mathcal{R}^+$), OAuth is the primary protocol, even if it is possible to transfer personal data and (links to) content over OpenID. Naturally, all relationships including actions use (and must use) OAuth.

3.4 Head-to-Head IDP Comparison

We have found that the top-three English speaking IDPs (Facebook, Twitter, and Google) are used differently by their RPs and that the usage is relatively independent of which other IDPs the RPs are using.

Table 4 breaks down the app rights for RPs using each of these three IDPs. Google is the only IDP with type $\mathcal{R}^-$ relationships. Most of these relationships are due to use of Google's OpenID-based API, which in comparison to Google's OAuth API typically share less information and does not allow actions.

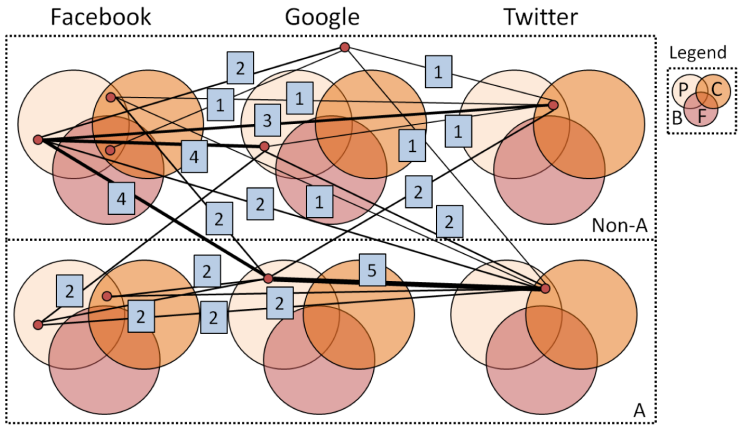


Fig. 3. Dependencies between app rights for the top-three English speaking IDPs. Here, the top-right legend shows the labeling of RP-IDP relationship types for each IDP, and link weights specify the number of RPs with such a relationship pair.

Table 4. Breakdown of risk types of the RP-IDP relationships associated with the top-three English speaking IDPs.

	Relationship type								
IDP	**Tot**	$\mathcal{R}^-$	$\mathcal{R}$	$\mathcal{R}^+$	$\mathcal{R}^{++}$	$\mathcal{R}_A$	$\mathcal{R}_A^+$	$\mathcal{R}_A^{++}$	**Unk**
Facebook	55	0	24	5	3	13	3	1	6
Twitter	15	0	0	4	0	0	11	0	0
Google	29	4	7	0	0	12	0	0	6

Overall, Google's mix of OpenID-based and OAuth-based relationships share less information (large fraction of $\mathcal{R}^-$, $\mathcal{R}$, and $\mathcal{R}_A$) than Twitter and Facebook (who also have many $\mathcal{R}^+$, $\mathcal{R}^{++}$, and $\mathcal{R}_A^+$ relationships). Furthermore, compared to Twitter and Facebook, Google allows more fine-grained personalization of app-rights agreements. The user is sometimes able to select which contacts (if any) to share information with (e.g., using Google's concept of "circles"). The most privacy preserving choice is, however, typically not selected by default.

Facebook is dominated by $\mathcal{R}$ and $\mathcal{R}_A$ relationships, and typically allows rich datasets to be imported to the RP. For Twitter, public messages and contacts are normally the only shared data (counted as content (C)); however, it should be noted that there are Twitter relations in which even private messages are shared with the RP. Twitter has the largest fraction of relationships with actions ($\mathcal{R}_A$ and $\mathcal{R}_A^+$). Twitter is particularly attractive for RPs wanting to perform actions on behalf of their users, as it provides an API that allows a wide of range of such actions to be performed and has a relatively active user population.

While we have not found any major statistical biases when comparing the fraction of relationship types observed when two IDPs (an IDP pairing) are used by the same RP vs. when they do not appear in such a pairing, we have found that there are some relatively common combinations. Figure 3 shows app-rights selections for IDP pairs consisting of Facebook/Google (19 pairs), Facebook/Twitter

(11 pairs), and Google/Twitter (12 pairs). For example, we note that RPs importing personal data (P) from Facebook, often do the same with Google (with or without actions). We also observe several cases where Google and Twitter are used together and both IDPs use actions (A) and importing personal (P) data (being classified as $\mathcal{R}_A$). In general, there is a bias for selecting to use actions (A) with one IDP, given that actions are used with the other IDP. For example, in 24 of 40 cases (60%) in which an RP-IDP relationship of a pairing uses action (A) the other relationship uses actions (A) too. In contrast, only in 16 of 44 cases (37%) in which the first RP-IDP relationship does not use actions (A), the other does. Using one-sided binomial hypothesis testing, these differences are significant with 98% confidence ($p = 0.015, z = 2.167$).

4 Multi-account Information

It is becoming increasingly common that users have accounts with multiple of the RP's IDPs. For example, in our original Soundcloud example (Figure 1), a user may have accounts with both Google and Facebook. In addition, a local RP account may be created either before connecting the account to one of the IDPs, or when first creating the account using one of the IDPs. The use of all these accounts and their relative dependencies can complicate the situation for the end user, potentially increasing privacy risks.

In this section we present the highlights from a set of targeted scenario-driven case studies that we have performed to analyze the interaction between the different accounts, as observed in the wild. (Due to space limitations detailed results and tables are omitted.) For this analysis, we performed tests for each pairing of the three most popular English-speaking IDPs: Facebook, Twitter, and Google. For each possible IDP pairing, we allowed both IDPs in the pair to be used first in a sequence of tests. The tests were also performed both with and without first creating local accounts at the RPs. For each test sequence, we recorded all information $I_{u(\alpha\to\gamma)}$ (of class B, P, C or F) that a user u agrees that the RP γ can import from IDP α, all information $I_{u(\gamma\to\alpha)}$ that user u agrees that the RP can post on the IDP (through actions (A)), all information $I_{u(u\to\gamma)}$ that the user manually inserts into its local profile, and the information $I_{u(p)}$ which ends up in the user profile.

4.1 Information Collision

When looking closer at the overlap between the information shared by the IDPs (i.e., $I_{u(\alpha\to\gamma)}$ and $I_{u(\beta\to\gamma)}$) with the RP, we observe that contact lists (26 of 42) are the most common overlap. As Twitter does not explicitly list email address, profiles picture, and names in their agreements or bundles, the overlaps for these categories are limited to Facebook+Google scenarios: 14 out of 42 cases for email addresses, 10 of 42 for profile pictures, and 10 of 42 for names. Having said this, we did observe cases where the profile picture and name were imported from Twitter without asking (or being listed in the app-rights agreement), suggesting

that these numbers only provide a lower bound of the potential information collisions.

As the shared information can be both conflicting and complementary, significant identity management complications can arise because of overlapping information. Yet, we have found little to suggest that the RPs provide users with identity/information management for these instances. In fact, even among the typically very small subset of information transferred to the user profile (Section 4.3) there often is an overlap. For example, regardless if there exists an initial local account or not, in 9 of 42 cases, at least some potentially conflicting information is imported to the user's RP profile from both IDPs.

4.2 Account Merging and Collisions

We next evaluate how well the RPs allow multiple accounts to be associated with a single user, and if the RPs allow multiple accounts to be merged. For merging to take place, the RP must allow the user to connect an IDP to an existing local account, or to connect a second IDP to an account that already have an IDP associated to it, such that both IDPs can be used to login to the RP.

Interestingly, we have found that both account merging and the information transferred between accounts often are highly dependent on the order in which accounts are added. Furthermore, in many cases the user is not able to merge accounts, or control if merging should take place. For example, out of the 42 (6x7) first-login cases when using a local account, only 10 cases resulted in optional merging and 11 in automatic merging. In 6 of the remaining cases, temporary accounts were created that did not have full functionality, in 7 cases the login failed altogether (typically due to collisions of email accounts between the original local account and that used at the IDP), and in the remaining cases a new account is created.

Similarly, out of the 2x42 (2x6x7) second-login cases with a second IDP, starting both with and without a local account, we observe few merging opportunities: 9 optional cases and 2 automatic cases, when there is no local account (12 and 10 cases when there is a local account). In total, a second IDP can be added (and merged) in 33 (11+22) of 84 cases.

Our results suggest that many RPs are not designing for multi-IDP scenarios, but that Facebook is doing the best job allowing for such relationships. The lack of multi-IDP support can have serious negative consequences as many of these IDPs are popular services with many users; increasing the chance that users have accounts with multiple IDPs. In the following, we take a closer look at information flow in the cases when two IDPs could be added.

4.3 Cross-IDP Information Leakage

Not only can information $I_{u(\alpha\rightarrow\gamma)}$ be shared from an IDP α to an RP γ, but in some cases the app-rights agreements with another IDP β may (intentionally or unintentionally) allow information to be moved from one IDP to another IDP (through the RP). This occurs when this agreement allows the RP γ to post some

subset of this information to IDP β. Looking at the overlap $I_{u(\alpha\to\gamma)} \cap I_{u(\gamma\to\beta)}$ we observed multiple cases where such cross-IDP sharing is possible. For example, six RPs allow personal (P) and/or content (C) from Facebook to be posted on Twitter, and five RPs allow basic (B) information from Facebook or Google to be transferred. We have also observed two RPs that have general posting rights on Facebook that allow transfer from Google, and two RPs that allow Facebook to transfer data from Twitter (although in this case Twitter would only transfer profile picture and name to the RPs). While these results show that all IDPs can be potential information sources and publishers of leaked information, in general, we see that Facebook allows the richest cross-IDP leakage and Twitter is the most likely publisher of cross-IDP leaked information.

5 Related Work

The third-party authentication landscape has gone from a situation with many OpenID enabled accounts but very few RPs supporting login with an IDP [13] to a landscape dominated by OAuth and Web 2.0 services that share data between sites. Limited work has been done on characterizing this emerging landscape.

In this paper, we leverage our RP-IDP relationship identifications and manual collection methodology [15], previously used to characterize and compare the relative biases (geographic and service-based) that RPs have in their IDP selection relative to the biases in the third-party content delivery landscape. In contrast, this paper looks closer at the information sharing and privacy risks associated with the observed RP-IDP relationships.

Sun et al. [14] have shown that there are misconceptions about how SSO works and the information transfers using OAuth, which leads to many users avoiding SSO. However, it can perhaps be argued that the users' fear is partially justified, as it has been shown that users become increasingly susceptible to phishing attacks as they get used to being redirected between sites and providing their credentials [5].

Others have proposed recommendation systems to help users make informed choices about what data to allow on different IDPs [11] and frameworks that provide users added control over how third-party applications (TPAs) can access data from an online social network (OSN) [4, 6]. Many of the insights in these works are directly applicable to ours, as TPAs play a similar role as our RPs and the OSN acts as IDP. Extending these frameworks to the general RP-IDP landscape, characterized here, provides interesting future work.

Many researchers have used the data available on the popular IDPs, sometimes generated with the help of RPs, to illustrate how data mining and other statistical methods can be used to determine potentially private information from public data (such as likes, twitter feeds, etc.) [3, 18, 10], or correlating user data from several websites to identify a user based on behavior [8]. This type of cross correlations is, of course, even easier if RPs and IDPs are linked. We see these examples as motivation for looking more closely into the increased cross-site sharing and potential information leakage associated with the RP-IDP

relationships. Cross-site leakage and the associated privacy risks have also been studied in the context of ad services and trackers [9].

Birrell and Schneider [2] present a privacy-driven taxonomy of the design choices in the third-party identity management landscape. Protocol related security problems that enable identity theft and blackmailing [17], and economic factors [5] have also been discussed in the literature.

6 Discussion and Conclusions

This paper characterizes the cross-site information sharing and privacy risks in the third-party identity management landscape. We show that OAuth is the dominant protocol, and OpenID is decreasing in usage. Not only is OAuth used by the most popular IDPs (e.g., Facebook, QQ, and Twitter) but it also enables sharing of much richer information with RPs.

We also observe high concentration in usage of a few popular IDPs, and that some IDPs are more frequently used in combination with others. The skew towards a few popular IDPs, which often allow RPs to act on behalf of the users, has privacy implications for the end users, regardless of whether the users choose to create an account with the RP or not. For example, a user with a compromised IDP account could easily be impersonated across all the RPs using a particular IDP, even if the user did not originally have an account with these RPs.

We then carefully classify and analyze the app-rights agreements of the most popular websites. Our classification of RP-IDP relationships is based on both the information that the RPs are allowed to import (e.g., basic, personal, generated content, or friend data) from the IDP, and if the RP is allowed to perform actions (e.g., create, update, or delete information) on behalf of the user on the IDP. Although we observe significant differences in the information leakage risks seen both across classes of RPs and across popular IDPs, we find multiple high-risk sites (e.g., RPs that both import private information and that are authorized to perform actions) among the top-200 websites for all website classes except Ads/CDN services. Such sites can easily become a source of information leakage.

Our multi-account case studies show that users are often asked to allow the RP to import more information from the IDP than is needed for the local user profile, and to enter redundant information. Furthermore, we find significant incompatibilities and inconsistencies in scenarios involving multiple IDPs. Often it is not possible to merge accounts with different IDPs, and the user can be stuck with a wide range of undesirable account situations. Clearly, many RPs are not designed to simply and securely use multiple IDPs.

We believe that more focus must be placed on multi-IDP scenarios when defining future policies for OpenID Connect (based on OAuth2), for example. Ideally, the user should remain in control of exactly what is being shared between the involved parties, and which information should be used and shared with each IDP. Similar to how Google+ (and to some extent Facebook) allows users to define circles, we believe that carefully designed protocols with added user control can be defined in this context. Already today, these IDPs allow some

degree of personalization in their RP-IDP permission agreements, so generalizing this concept to the context of multiple IDPs could be one possible approach. Future work will include the definition and evaluation of such policies.

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. E. Birrell and F. B. Schneider. Federated identity management systems: A privacy-based characterization. *IEEE Security & Privacy*, 11(5):36–48, Sep/Oct. 2013.
3. A. Chaabane, G. Acs, and M. A. Kaafar. You are what you like! information leakage through users' interests. In *Proc. NDSS*, Feb. 2012.
4. Y. Cheng, J. Park, and R. Sandhu. Preserving user privacy from third-party applications in online social networks. In *Proc. WWW*, May 2013.
5. 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.
6. A. Felt and D. Evans. Privacy protection for social networking APIs. In *Proc. W2SP*, May 2008.
7. P. Gill, M. Arlitt, N. Carlsson, A. Mahanti, and C. Williamson. Characterizing organizational use of web-based services: Methodology, challenges, observations, and insights. *ACM Trans. on the Web*, 5(4):19:1–19:23, Oct. 2011.
8. O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting innocuous activity for correlating users across sites. In *Proc. WWW*, May 2013.
9. D. Malandrino, A. Petta, V. Scarano, L. Serra, R. Spinelli, and B. Krishnamurthy. Privacy awareness about information leakage: Who knows what about me? In *Proc. ACM WPES*, Nov. 2013.
10. M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proc. ACM SIGKDD*, Aug. 2011.
11. M. Shehab, S. Marouf, and C. Hudel. Roauth: Recommendation based open authorization. In *Proc. SOUPS*, July 2011.
12. 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.
13. 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.
14. S.-T. Sun, E. Pospisil, I. Muslukhov, N. Dindar, K. Hawkey, and K. Beznosov. Investigating user's perspective of web single sign-on: Conceptual gaps, alternative design and acceptance model. *ACM Trans. on Internet Technology*, 13(1):2:1–2:35, Nov. 2013.
15. A. Vapen, N. Carlsson, A. Mahanti, and N. Shahmehri. Third-party identity management usage on the web. In *Proc. PAM*, Mar. 2014.
16. A. Vapen, N. Carlsson, A. Mahanti, and N. Shahmehri. Information sharing and user privacy in the third-party identity management landscape. In *Proc. ACM CODASPY*, Mar. 2015.
17. 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.
18. R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. In *Proc. ACM SIGKDD*, Aug. 2013.