# Information Sharing and User Privacy in 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

Third-party identity management services enable cross-site information sharing, making Web access seamless but also raise significant privacy implications for the users. Using a combination of manual analysis of identified third-party identity management relationships and targeted case studies we capture how the protocol usage and third-party selection is changing, profile what information is requested to be shared (and actions to be performed) between websites, and identify privacy issues and practical problems that occur when using multiple accounts (associated with these services). The study highlights differences in the privacy leakage risks associated with different classes of websites, and shows that the use of multiple third-party websites, in many cases, can cause the user to lose (at least) partial control over which information is shared/posted on their behalf.

## Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

## Keywords

Third-party Identity Management; Cross-site Information Sharing; User Privacy

## 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 information can also be used to provide personalized and customized user experiences that add value for their users. Therefore, many other websites are partnering with these companies, often using third-party single sign-on (SSO) [1, 2] services provided by these and other popular websites.


*CODASPY'15,* March 2–4, 2015, San Antonio, Texas, USA.
ACM 978-1-4503-3191-3/15/03.
http://dx.doi.org/10.1145/2699026.2699131.

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. In this scenario, Soundcloud is referred to as a relying party (RP) and Facebook/Google is referred to as a third-party identity provider (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-right agreement between the user and the RP, which (i) gives permission to the RP to read 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 also place great responsibility on the RPs, and can raise significant privacy concerns for users.

The paper makes three primary contributions. First, we present a high-level characterization of the protocol and IDP usage observed in the wild (Section 2). Second, we characterize the cross-site information sharing and authorized app-rights associated with the most popular IDPs (Section 3). Third, 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). For a complete description of our methodology, contributions, and a discussion of the above results we refer to our full paper [3]. Here, we briefly describe some high-level results.

## 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 user's IDP account (even though the user must not remember any such account information later), allowing information sharing between the RP and IDP. Local RP accounts are optional with OpenID.

For our analysis, we primarily focus on all RP-IDP relationships that we have manually identified on the 200 most popular websites on the Web, 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 [4].

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 di-

rectly 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 other IDPs. 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 IDPs ranked in the top-5. Facebook+Twitter is the most popular pairing with 37% (125 of 335) of all IDP pairs, Chinese QQ+Sina placing second (19%), and Facebook+Google third (12%).

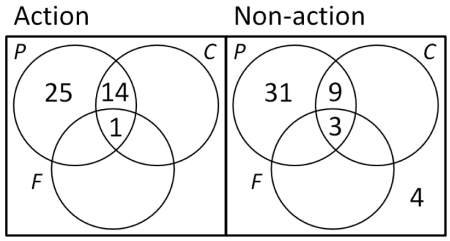


**Figure 1: Number of RP-IDP relationships of different app-right types in the top-200 dataset.**

**Table 1: Risk types identified in dataset.**

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

# 3. APP RIGHTS AND INFORMATION FLOWS

We carefully recorded the app-right agreements for the RP-IDP relationships identified in the manual top-200 dataset. The app-right agreements reveal (i) the information that the RP *will obtain* from the IDP, and (ii) the actions the RP *will be allowed* to perform on the IDP, on behalf of the user.

## 3.1 Classification of Information

When analyzing the APIs of the three major IDPs (Facebook, Twitter and Google) and the actual app-right usage in our datasets, we have identified five different types of app rights, each with their own privacy implications. The first four classes (B, P, C, F) capture data transferred from the IDP to the RP. Class A includes 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. This class includes unique identifiers (e.g., user name, id, or email address) to identify existing accounts, age range, language, and public user profile 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.
- **Created content (C):** This class contains content directly or indirectly created by the user (e.g., images, likes, and check-in history).
- **Friends' information (F):** This class consists of data of other potentially non-consenting users (e.g., friends).
- **Actions taken on behalf of the user (A):** This final class includes the right for the RP to perform *actions* on behalf of the user on the IDP. This include, for example, the right to post information about the user's actions on the RP (e.g., sharing music the user has listened to) on the user's IDP timeline.

## 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. This simplifies the agreements, but reduces the 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 1 summarizes all the observed app-right 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. The small number of pure B relationships (4), suggests that there appear to be an expectation of trust in the RPs, beyond what the user typically 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-right agreements that does not request actions to be performed, only 12 (9+3) also request access to content (C).

We note that within each of the two boxes there is a clear ordering in risk types observed. 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. Motivated by these observations, we identify 8 semi-ordered risk classes. Table 1 summarizes the observed classes. 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. Among the classes with at least 10 RPs, News sites and File sharing sites are the most frequent users of actions (risk types $\mathcal{A}$ and $\mathcal{A}^+$), with 55% and 50% of their relationships including actions, respectively. Also Video sharing (67%) and Tech (63%) sites has large fraction of relationships that include action (A) permissions. The high action (A) usage is likely an effect of these sites often wanting to promote contents to friends of the user. While we express privacy concerns regarding $\mathcal{A}^+$ relationships, these sites would in

**Table 2: Breakdown of relationship types for the top-three English speaking IDPs.**

| | Relationship type | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **IDP** | **Tot** | $\overline{\mathcal{A}^-}$ | $\overline{\mathcal{A}}$ | $\overline{\mathcal{A}^+}$ | $\overline{\mathcal{A}^{++}}$ | $\mathcal{A}$ | $\mathcal{A}^+$ | $\mathcal{A}^{++}$ | **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 |

fact desire that the information that their content are being read/watched to propagate across many sites.

Relationships including actions are primarily associated with RPs that have many IDPs. For example, while RPs with one IDP use actions in 33% of their relationships (all using Facebook as their only IDP), RPs with multiple IDPs use actions in 48-53% of their relationships. As with our discussion about News and File sharing sites, the many IDPs of these RPs increases the risk for cross-site leakage.

The most restrictive type ($\overline{\mathcal{A}^-}$) includes only OpenID. Even if OpenID allows some data transfer, OAuth is the primary protocol for content sharing without actions ($\overline{\mathcal{A}^+}$). Naturally, all relationships including actions use OAuth.

## 3.4 Head-to-Head IDP Comparison

The top-three English speaking IDPs are used relatively differently by their RPs and the usage is relatively independent of which other IDPs the RPs are using.

Table 2 breaks down the app rights for RPs using each of these three IDPs. Google is the only IDP with type $\overline{\mathcal{A}^-}$ relationships. Google's mix of OpenID-based and OAuth-based relationships share less information than Facebook. Facebook typically allows rich datasets to be imported to the RP. For Twitter, public messages and contacts are normally the only shared data. Twitter is particularly attractive for RPs wanting to perform actions on behalf of their users. 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 import personal (P) data (classified as type $\mathcal{A}$). In general, there is a bias for selecting to use actions (A) with one IDP, given that actions are used with the other IDP.

# 4. MULTI-ACCOUNT INFORMATION

It is becoming increasingly common that users have accounts with multiple of the RP's IDPs. 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.

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 $I_{u(\alpha\rightarrow\gamma)}$ (of type B, P, C or F) that a user $u$ agrees that the RP $\gamma$ can import from IDP $\alpha$, all information $I_{u(\gamma\rightarrow\alpha)}$ that user $u$ agrees that the RP can post on the IDP (through actions (A)), all information $I_{u(u\rightarrow\gamma)}$ that the user manually inserts into its local profile, and the information $I_{u(p)}$ which ends up in the user profile.

**Information collision:** Significant identity management complications can arise because of overlapping information shared by the IDPs (i.e., $I_{u(\alpha\rightarrow\gamma)}$ and $I_{u(\beta\rightarrow\gamma)}$) and the RP. We find that contact lists (26 of 42) are the most common overlap, and that 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.

**Account merging and collisions:** 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.

**Cross-IDP information leakage** Looking at the overlap $I_{u(\alpha\rightarrow\gamma)}\cap I_{u(\gamma\rightarrow\beta)}$ we observed multiple cases where cross-IDP sharing is possible, allowing information to be moved from one IDP to another IDP (via the RP). 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).

# 5. CONCLUSIONS

Using targeted case studies on both manually and automatically identified RP-IDP relationships, this paper characterizes the cross-site information sharing and privacy risks in the third-party identity management landscape. We observe significant differences in the information leakage risks seen both across classes of RPs and across popular IDPs. Yet, for all website classes except Ads/cdn services, we find multiple high-risk sites among the top-200 websites. This includes RPs that both import private information and that are authorized to perform actions on the IDP. Furthermore, we find significant incompatibilities and inconsistencies in scenarios involving multiple IDPs. Clearly, many RPs are not designed to simply and securely use multiple IDPs. 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.

# 6. REFERENCES


[1] 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.

[2] 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.

[3] A. Vapen, N. Carlsson, A. Mahanti, and N. Shahmehri. Information sharing and user privacy in the third-party identity management landscape. Technical report, 2014.

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