Taking Data Exposure into Account: How Does It Affect the Choice of Sign-in Accounts?

Shahar Ronen
MIT Media Lab
sronen@media.mit.edu

Oriana Riva
Microsoft Research
oriana@microsoft.com

Maritza Johnson
U. of California, Berkeley
maritzaj@cs.berkeley.edu

Donald Thompson
Microsoft Research
donthom@microsoft.com

ABSTRACT
Online services collect personal data from their users, sometimes with no clear need. We studied how users sign-in to web sites using federated IDs, and found that most survey respondents were not aware of the data they expose. However, when presented with the tradeoffs behind each sign-in option, respondents reported a willingness to change how they sign-in to reduce their data exposure or, in fewer cases, to increase it to receive more benefits from the service. Our findings suggest that data exposure is a concern for users, and that there is a need for finding clearer ways for communicating it for each sign-in option.

Author Keywords
Online services; sign in; federated identity; data exposure.

ACM Classification Keywords
H.5.m. [Information Interfaces and Presentation]: Misc.

INTRODUCTION
Many web sites require users to register to enjoy additional benefits: access to premium content (Bloomberg), storage and sharing of media (Flickr), a personalized experience (Monster, The Huffington Post), etc. Registration is the establishment of a relationship with the site using a permanent identity, which can be created on the site or provided by a third party (*federated identity* [4, 9, 11]), such as an e-mail provider (Google, Microsoft, Yahoo!) or a social networking service (Facebook, Twitter). Users can register with Flickr, for instance, using a Facebook, Google, or Yahoo! account. Yet, in addition to benefits, registering with a web site has a cost: personal data are collected by the site and used for its own benefit [1,13]. This is especially true when signing in with federated IDs, which may hold many personal data about users. For example, by registering with Flickr using their Facebook account, users grant Flickr access to information about their Facebook friends.

Some sites support registration through multiple federated IDs, with different data exposure and benefits tradeoffs (Table 1 and 2). Federated ID providers often inform users of the data they expose to a particular site (yet not necessarily in a clear way), but rarely let users specify upon registration the types of data made available to the site. Even when users are allowed to decide what to share, they have a hard time choosing, as most sites do not explain why a certain type of data is required and what benefit is provided in return.

In this paper, we present a study that explores how well users are aware of personal data transferred from their federated accounts to sites upon sign-in, and how their sign-in choices change after they are presented with more information about the personal data passed to the site and the benefits received in return. We surveyed 575 people over a two-week period about their sign-in preferences for real and invented sites. We recruited registered and non-registered users to participate using Mechanical Turk. Most of our participants were not aware of the types of personal data they expose to sites upon sign-in, but when given better notice many changed their sign-in choices to reduce data exposure or, in fewer cases, to increase exposure to enable features they perceived as useful. The frequency of both these changes suggests that users do not understand the tradeoffs associated with different sign-in options.

Personal data type	Mons. acct.	Yahoo!	FB
Name and e-mail address	X	X	X
Your picture			X
Gender and birthdate			X
Bio or description		X	X
Interests		X	X
Contacts / friends			X
Education and work history	X		X
Location	X		X
Contacts' info (desc., interests, etc.)			X

Table 1: Data sharing table for Monster, showing for each sign-in option the personal data passed to the site.

Feature	Mons.	Yahoo!	FB
Create your professional profile and notify you about relevant job opportunities	X	X	X
Automatically fill in your professional profile with relevant information			X
Use your friends' education and work info to show you where you have inside connections			X

Table 2: Benefits table for Monster, showing for each sign-in option the features it enables.

RELATED WORK
User misconceptions of data exposure have been studied before, but not in the context of federated IDs. King et al. [8] found that many users were not able to identify the data types exposed by Facebook to third-party apps. Facebook now lists transferred data in the app installation consent

CHI 2013, April 27–May 2, 2013, Paris, France.

form, but these could be misleading [3]. Also, studies on consent forms have shown that users do not understand what they consent to [5, 6]. The use of federated IDs poses additional challenges. First, there are more options to choose from and each may expose different data. Second, users do not understand how federated IDs work: 70% of respondents to Sun et al.'s study [12] thought that the identity provider passed their username and password to the website. Our study goes beyond usernames and passwords and investigates transfer of different types of personal data.

Kelley et al. [7] found that the speed and accuracy of decisions about privacy improved when users were presented a standardized table showing the personal data stored by the site. We use a similar format for communicating to users the data disclosed, but add a table showing the benefits for each option to provide participants complete information about the tradeoffs incurred. This allows us to measure how a better understanding of data exposure affected users' choice of a sign-in account among multiple options.

METHODOLOGY AND SURVEY STRUCTURE

We started by asking our participants how many accounts they have with the following providers: Facebook (FB), Google (G), Microsoft (MS), and Yahoo! (Y), and which of the following data types they believe are stored in each account: basic contact details (phone, e-mail, address, etc.), picture, work and education history, bio/description, location, interests, e-mails, chat history, contacts, events, documents, photos, status updates, contacts' info (description, interests, etc.), and content generated by contacts (documents, status updates, etc.). Then, we presented participants with one of the following sites: Monster, Flickr, The Huffington Post, and Outgo!ng. We selected these sites through a pre-survey, asking 190 Turkers which of 38 sites they were registered with. Flickr ranked first (37% of participants), followed by Monster (15%). We also chose The Huffington Post (6%) because of its wide variety of registration options. Outgo!ng is an invented site, described to participants as "a site for scheduling recreational activities with friends. Given a time frame, Outgo!ng suggests available friends and nearby activities that match your mutual interests".

Participants received one of two questionnaires depending on whether they were registered with the site. Registered users were asked to select from a list the account they used to sign in. The list was populated with accounts supported by the site and also previously specified by the participant. Sign-in options for each site were as follows. HuffPost: FB, G, MS, Y, and a proprietary account that can be created on the site. Flickr: FB, G, Y. Monster: FB, Y, and a Monster account. Outgo!ng: FB, G, MS, Y, and an Outgo!ng account. Finally, we asked participants to select from the list of data types the ones they thought were transferred from the account to the site. Non-registered participants were presented a description of the site, and asked to select from a list the account they would use to register. The list was similar to the list presented to registered users, with the additional option of creating a new account with one of the identity providers supported by the site.

Then, we presented both groups with two tables: one listed, for each sign-in option, the data types actually transferred from the account to the site, and the other listed the benefits provided by each option (Table 1 and 2). We then asked participants whether this information made them reconsider their sign-in choice. If so, we asked to select the option that best described why: "I am more comfortable transferring to the site the personal information stored in this account" (interpreted as a concern about data exposure), "I realized I could get a better experience by choosing this account" (desire for better benefits), and "Other (please specify)".

Data Collection

We ran our survey using Amazon's Mechanical Turk over two weeks in August 2012. We aimed for replies from 200 people for each of the real sites (100 registered users) and 50 for Outgo!ng but did not meet the goal (Table 3). We limited participation to US residents with a HIT approval of over 95%, and paid on average $0.45 per survey. Participants were free to choose a HIT (and thus a site) but were limited to submitting one survey. Registered users had to prove they were registered to the site by signing in and copying unique text (e.g., menu options). We received 575 valid responses split almost equally across genders (288 M, 285 F, 2 undisclosed) and from a range of ages (52% aged 18-27, 27% aged 28-37, 21% aged 38+). 86% of our respondents had at least one Facebook account; Google (80%), Yahoo! (64%), and Microsoft (38%) followed.

Site	Registered	Non-registered	Total
Outgo!ng (invented)	N/A	46	**46**
Flickr	61	113	**174**
Huffington Post	49	123	**172**
Monster	67	116	**183**
Total	***177***	***398***	***575***

Table 3: User registration by site.

Site	FB	G	MS	Y	Site acct.	New acct.	Other	Total
Outg NR	12	6	2	6	-	20	-	**46**
Flickr R	2	13	-	40	-	-	6	**61**
Flickr NR	42	35	-	21	-	14	1	**113**
HPost R	9	12	2	7	18	-	1	**49**
HPost NR	18	27	13	20	-	44	1	**123**
Mons R	5	-	-	4	55	-	3	**67**
Mons NR	10	-	-	22	-	79	5	**116**
Total	***98***	***93***	***17***	***120***	***73***	***157***	***17***	***575***

Table 4: Distribution of sign-in accounts by site and registration status: registered (R) and not registered (NR). Numbers for account providers indicate existing accounts; *New acct* is a newly created account with any provider.

RESULTS AND ANALYSIS

Despite the concerns rightfully raised by Sun et al. [12] about user adoption of single sign-on solutions, we found that participants were generally open to using federated accounts when given the option (Table 4). The share of registered users who signed in using federated accounts changed drastically from 13% for Monster to 61% for HuffPost (Flickr supports federated sign-in only). Unregistered users were more open to federated IDs, with

45%, 81%, and 78% choosing a federated account to sign in to Monster, HuffPost, and Outgo!ng, respectively.

Awareness of Personal Data

We compared the number of data types participants (registered and non-registered) thought were associated with each account (

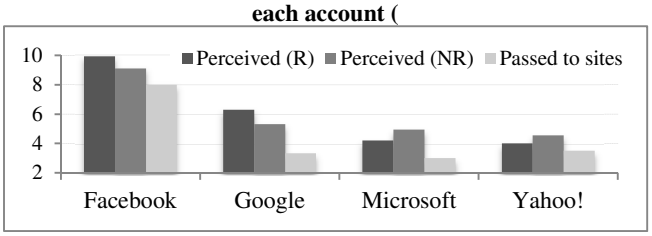

Figure 1). Facebook came first with an average of 9.37 data types, followed by Google (5.66), Microsoft (4.71), and Yahoo! (4.32). This indicates that respondents perceive their Facebook accounts as containing more personal data than other providers'. One explanation is that Facebook is a social network whereas the other providers are mostly seen as e-mail services; Google may rank higher than Microsoft and Yahoo! because of its social network, Google+. User perception seems to be correct, as our sites had more data types transferred from Facebook accounts (an average of 8 types per user) than from Yahoo!, Google, or Microsoft, (3.5, 3.3, and 3 types, respectively).

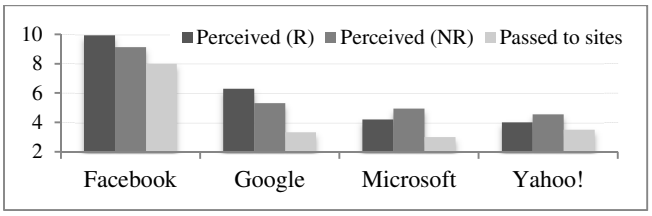

Figure 1: Average number of data types per account. First two bars show what (registered and unregistered) users perceived; the last bar shows the actual number of types passed to sites.

Despite their good intuition about the variety of data stored, participants were not able to correctly identify the data types transferred to the sites. We asked registered users who sign-in with federated accounts (n=94) to select the data types they thought were passed to the site. We compared their answers with the actual data types passed to each site, measuring relevance of responses using *precision* and *recall*. We define precision (P) as the ratio between the number of data types a user named correctly and the number of data types the user named, and recall (R) as the ratio between the number of data types a user named correctly and the number of data types that were actually transferred to the site. For example, if a user specifies that "basic info" and "location" are transferred to a site and the actual data transferred are "basic info", "picture" and "contacts", P=0.5 and R=0.33. The average across all 94 users was P=0.48, and R=0.64 (standard deviation of 0.37 and 0.39, respectively). This means that participants only identified two-thirds of the data types transferred, and that about half the types they guessed were incorrect, showing a rather limited awareness of data types transferred.

Account Reconsideration

Presented with the data sharing and benefit tables, 255 participants (44%) said they would consider changing the sign-in account they had originally chosen. The same rate applied to registered and non-registered users, and we found similar rates (43%-46%) across all sites. However, only 196 of the 255 (34% of all participants) actually named a different account when asked to select the new sign-in account. We refer to them in the next analysis.

Site	Data exposure	Better benefits	Not resolved	Total
Flickr	43 (21.9%)	18 (9.2%)		**61 (31.1%)**
HuffPost	33 (16.8%)	23 (11.7%)	1 (0.5%)	**57 (29%)**
Monster	36 (18.4%)	27 (13.8%)		**63 (32.2%)**
Outgo!ng	6 (3.1%)	9 (4.6%)		**15 (7.7%)**
Total	***118 (60.2%)***	***77 (39.3%)***	***1 (0.5%)***	***196 (100%)***

Table 5: Reasons for change given by "account changers". "Other" responses were resolved to "data exposure" (16) and "better benefits" (1), but one could not be resolved to either.

Most "account changers" indicated data exposure as the main reason for change (118 participants), while fewer (77) were attracted by the option of more benefits (Table 5). The difference is especially noticeable in the case of Flickr, where the "data-concerned" outnumber the "benefit-concerned" 2.4 to 1. These were mostly Facebook users: 24 (55%) of 44 participants (2 registered, 42 unregistered) who originally chose Facebook to sign in to Flickr changed their decision. A majority of these participants (19, including the two registered users) said they would do so because of a concern about the data Facebook transfers to Flickr; only five mentioned better benefits as the reason for change. Indeed, signing in to Flickr using Google or Yahoo! only transfers a user's e-mail to the site, while a Facebook sign-in transfers e-mail address, picture, gender and birthplace, interests, friends, work and education history, events, status updates, details of friends, and status updates by friends.

	Data exposure	Better benefits	All users
Flickr	-4.19 (43)	+2.22 (18)	**-2.30 (61)**
HuffPost	-1.82 (33)	-0.35 (23)	**-1.21 (57)**
Monster	-0.28 (36)	+1.11 (27)	**+0.32 (63)**
Outgo!ng	-3.50 (6)	-0.44 (9)	**-1.67 (15)**
Total	***-2.30 (118)***	***+0.75 (77)***	***-1.09 (195)***

Table 6: Change in the average number of data types an "account changer" potentially exposed to a site before and after being presented with the tradeoff tables, by reason given for change. Only resolved reasons are shown (n=195).

Change in Data Exposure

We found that participants concerned about their data exposure were able to take effective measures against it. Although the average number of data types potentially exposed to sites decreased for all participants, those who changed account providers because of data exposure exhibited a larger reduction than the benefit-minded participants (-2.30 vs. +0.75, Table 6). It seems that when participants were presented with clearer tradeoffs between data exposure and benefits, they were able to immediately understand them and act accordingly. Decrease in exposure was manifested across all data types and almost all sites (Figure 2). The only site for which data exposure increased

was Monster: the 27% increase in the exposure of profile picture, contacts, and contacts' info is the result of a 6% increase in sign-in to Monster using Facebook (Figure 3). Monster uses the additional data from Facebook to identify "inside connections" in potential work places, and users seem to like it: "better benefits" was the reason given by 9 of the 12 participants who switched to a Facebook account.

	Profile picture	History	Bio	Location	Interests	E-mail content	Contacts	Status updates	Contacts' info	Contacts' content	Events
Flickr	-32%	-32%	-32%	-32%	-32%	N/A	-32%	-32%	-32%	-32%	-32%
Huff. Post	-21%	N/A	-24%	-19%	-24%	N/A	-19%	-30%	N/A	N/A	N/A
Monster	27%	3%	0%	3%	0%	N/A	27%	N/A	27%	N/A	N/A
Outgo!ng	N/A	N/A	N/A	-17%	-15%	-17%	-15%	-13%	-33%	-33%	0%
Total	-19%	-5%	-19%	-7%	-19%	-17%	-18%	-26%	-20%	-32%	-24%

Figure 2: Increase or decrease of "account changer" users per data type exposed to our four sites, after being presented with the data sharing and benefit tradeoff tables (n=196).

A simple measure to limit data exposure is the creation of a new account. A new account holds little or no personal information compared to an existing one, and participants seem to understand this: 42% of "account changers" switched from existing accounts to newly created accounts after they were shown the data sharing and benefit tradeoff tables (see Figure 3). While this trend was found across all sites, the abandoned accounts changed from site to site. For Flickr and Outgo!ng, Facebook exhibited the strongest decrease. For Monster it was the site account that suffered the most, whereas HuffPost exhibited a relatively equal churn rate across all existing accounts.

	Facebook	Google	Microsoft	Yahoo!	Site acct.	New acct.	Other
Flickr	-23%	3%	N/A	-11%	N/A	38%	-7%
Huff. Post	-9%	-7%	-7%	-14%	-11%	47%	0%
Monster	6%	N/A	N/A	-6%	-40%	46%	-6%
Outgo!ng	-27%	7%	-7%	0%	0%	27%	0%
All sites	-10%	-1%	-3%	-10%	-16%	42%	-4%

Figure 3: Increase of decrease of "account changer" users per sign-in account to our four sites, after being presented with the data sharing and benefit tradeoff tables (n=196).

LIMITATIONS

The design of our study allowed us to evaluate users' understanding of data exposure. The results on participant's willingness and reasons to change sign-in methods are subject to limitations. First, we collected self-reported data, which may not correlate to real changes. Second, while the word "privacy" was not mentioned [1, 2], references to data transfer may have increased participants' concern over data. Third, some participants may have felt that we expected them to change their behavior after reading the tradeoff tables and answered accordingly. Also, to facilitate analysis, we limited reasons for account change to "data exposure", "benefits" and "other", which may have biased participants towards these options. Finally, the use of Mechanical Turk may have introduced a selection bias.

CONCLUDING REMARKS

We studied user sign-in patterns to real sites, using federated and proprietary accounts, and found that the data awareness of most participants was lacking and they could not accurately identify which of their data were transferred from an account to the site. However, given a clear description of their data exposure and benefit tradeoff, participants reported a willingness to change sign-in methods to reduce their data exposure or to get more benefits; this applies to previously registered users and non-users. Our findings suggest that the consent forms currently used by online services do not allow users to make informed choices, and that explaining what data are transferred to a site and why would benefit users. Follow-up studies could evaluate user turnover more accurately, e.g., with participants who were not presented with detailed tradeoff tables as a control. Topics for future study include the effect of trust relationship with sites and account providers on the choice of sign-in account (following [12]), and the efficient presentation of tradeoff tables [7].

REFERENCES

1. Barkuus, L. The Mismeasurement of privacy: Using contextual integrity to reconsider privacy in HCI. In *CHI* '12, ACM (2012), 367-376.
2. Braunstein, A. et al. Indirect content privacy surveys: measuring privacy without asking about it. In *SOUPS* '11, ACM (2011).
3. Charkam, A. 5 design tricks Facebook uses to affect your privacy decisions. *Tech Crunch* (Aug 25, 2012). Accessed Sep. 5, 2012. http://techcrunch.com/2012/08/25/5-design-tricks-facebook-uses-to-affect-your-privacy-decisions/
4. Dhamija, R., and Dusseault, L. The seven flaws of identity management: Usability and security challenges. *IEEE Security & Privacy* 6, 2 (2008), 24-29.
5. Good, N. et al. Stopping spyware at the gate: A user study of privacy, notice, and spyware. *SOUPS* '05, ACM (2005), 43-52.
6. Grossklags, J., and Good, N. Empirical studies on software notices to inform policy makers and usability designers. In *USEC* '07, LNCS, Springer (2007).
7. Kelley, P. G., et al. Standardizing privacy notices: an online study of the nutrition label approach. In *CHI* '10. ACM (2010), 1573-1582.
8. King, J., et al. Privacy: is there an app for that?. In *SOUPS* '11, ACM (2011).
9. Maler, E., and Reed, D. The Venn of identity: Options and issues in federated identity management. *IEEE Security & Privacy* 6, 2 (2008), 16-23.
10. Shehab, M., et al. 2011. ROAuth: recommendation based open authorization. In *SOUPS* '11, ACM (2011).
11. Shim, S. S. Y. et al. Federated identity management. *Computer 38*, 12 (2005), 120-122.
12. Sun, S.-T. et al. What makes users refuse web single sign-on? An empirical investigation of OpenID. In *SOUPS* '11, ACM (2011).
13. What they know. *Wall Street Journal*. http://blogs.wsj.com/wtk.