

## Protecting Privacy on Social Media: Is Consumer Privacy Self-Management Sufficient?

Yaqoub Alsarkal  
George Washington University  
[alsarkal@gwmail.gwu.edu](mailto:alsarkal@gwmail.gwu.edu)

Nan Zhang  
American University  
[nzhang@american.edu](mailto:nzhang@american.edu)

Heng Xu  
American University  
[xu@american.edu](mailto:xu@american.edu)

### Abstract

*Among the existing solutions for protecting privacy on social media, a popular doctrine is privacy self-management, which asks users to directly control the sharing of their information through privacy settings. While most existing research focuses on whether a user makes informed and rational decisions on privacy settings, we address a novel yet important question of whether these settings are indeed effective in practice. Specifically, we conduct an observational study on the effect of the most prominent privacy setting on Twitter, the protected mode. Our results show that, even after setting an account to protected, real-world account owners still have private information continuously disclosed, mostly through tweets posted by the owner's connections. This illustrates a fundamental limit of privacy self-management: its inability to control the peer-disclosure of privacy by an individual's friends.*

*Our results also point to a potential remedy: A comparative study before vs after an account became protected shows a substantial decrease of peer-disclosure in posts where the other users proactively mention the protected user, but no significant change when the other users are reacting to the protected user's posts. In addition, peer-disclosure through explicit specification, such as the direct mentioning of a user's location, decreases sharply, but no significant change occurs for implicit inference, such as the disclosure of birthday through the date of a "happy birthday" message. The design implication here is that online social networks should provide support alerting users of potential peer-disclosure through implicit inference, especially when a user is reacting to the activities of a user in the protected mode.*

### 1. Introduction

Online social networks (OSNs), like Facebook and Twitter, enable highly interactive communications for online users but also introduce significant privacy risks

through the potential exposure of users' identity and personal information. As a result, extensive research has been done on how users view and protect their privacy in OSNs [31].

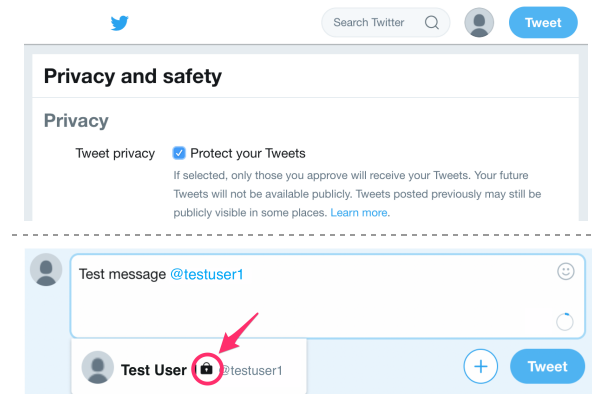
Among the existing solutions, *privacy self-management* [30] has been a popular choice. It comprises privacy-enhancing technologies and tools that allow users to protect their own privacy by directly controlling access to their personal data or shared information. Twitter's "Protected Mode" is one such example. After switching to protected mode, a user's past and future tweets, along with replies to these tweets, will be removed from the public view and become visible only to followers approved by the user.

Given its popularity, the effectiveness of privacy self-management, especially whether users are able to make informed and rational choices, has been debated for a long time, leading to many recent advances aiming at helping users with making better decisions [3, 8]. Notably missing, however, is a careful study of whether the choices themselves - as offered by a privacy self-management scheme - truly provide the protections that their descriptions would imply. For example, when Twitter boasts "*protect your tweets*" as the most prominent option under "Twitter Privacy", does it truly imply that a user's privacy can be protected once he/she starts protecting tweets? This is an important question because if none of the choices properly protects privacy, then self-management will not be effective no matter how user-friendly the privacy-setting interface is or how much decision support the user gets for making an informed decision.

This paper aims to address the gap in the literature by assessing the real-world effectiveness of privacy self-management choices. Specifically, we study the effectiveness of "protected mode" on Twitter, given its role as the most prominent privacy setting (see Figure 1) for an extremely popular OSN.

The first research question we study is whether private information about a Twitter user is still disclosed even after the user switches to protected mode. Unfortunately, recent behavioral research hint at a negative answer: Given the highly interactive nature

of interpersonal communications in OSNs, there has been a growing recognition of conceptualizing privacy as interdependent phenomenon between an individual and online contacts [5, 9]. These studies found while individuals are free to decide what information they disclose, they often have control over what others disclose about them. In the case of Twitter, while users can protect their own tweets through the privacy setting, they cannot prevent others from mentioning them (i.e., @UserA) in a public tweet and “peer-disclosing” certain private information about them.



**Figure 1.** Twitter protected-mode: top shows its prominence in privacy settings; bottom shows a lock icon when mentioning a protected user.

While the intuition behind this “*privacy peer-disclosure*” is straightforward - and empirical evidence of its existence on Twitter has been reported [4] for non-protected Twitter users - this paper reports the first evidence showing that, even after a Twitter user elects to protect all tweets, substantial private information, including photos, protected tweets, and sensitive information such as date of birth and home address, are still *continuously* disclosed to the public through peer-disclosure. Such peer-disclosure takes place in a variety of forms, e.g., the mentioning of a user (“@”) in a tweet, a well-known “off-the-book” trick (“dot-start”) of replying publicly to a protected tweet that is otherwise prohibited (by the Twitter system design) from being retweeted or quoted, etc.

After identifying the privacy leakage under protected mode, we looked for potential remedies by digging into *how* peer-disclosures occur for protected users. Note that Twitter prominently displays the protected status of a user when he/she is mentioned in a tweet or when his/her tweet is replied (see Figure 1)<sup>1</sup> - leading to the intriguing question of whether others

<sup>1</sup> Of course, we have no way of knowing whether others indeed see or understand the icon – an issue we leave for future studies.

will consider a user’s desire to remain private when publicly disclosing information about the user.

To this end, we compared the distributions of multiple types of peer-disclosures *before* and *after* a user set his/her tweets to protected. Interestingly, the answer to the above question depends on the form of disclosure, for which we introduce a 2×2 typology:

- 1) whether the peer user is *proactively mentioning* or *reactively interacting* with the protected user. The former occurs when the peer is initiating an activity that mentions (e.g., “@” in Twitter) the protected user; while the latter is incurred by the peer user reacting (e.g., replying) to an activity initiated by the protected user.
- 2) whether the disclosure is by *explicit specification* or *implicit inference*. The former represents activities that directly state the protected user’s information (e.g., spelling out the user’s full name), while the latter captures a peer’s activities that do not directly state such information, yet allow a third party to infer certain information about the protected user, e.g., gender through a “girl’s night out” message.

Equipped with this typology, we have a surprising finding that, once a user switches to the protected mode, there is a significant decrease of peer-disclosure by *proactive mentioning* and *explicit specification*, but not by *reactive interactions* and *implicit inference*, which remain unchanged. This suggests an interesting possibility that, while Twitter users do adjust their peer-disclosure behavior based on the observed “protected” status of others, they might not see the necessity of adjustment when the interaction was initiated by the protected users; and they might not be aware of the peer-disclosure when it happens through implicit inferencing. This possibility, in turn, suggests a remedy: alerting a user of potentially disclosing a protected user’s information, especially when the user is reactively interacting with the protected user and/or the disclosure is through implicit inference. We leave the study of this remedy to future work.

The contributions of this paper are summarized as follows. First, it provides empirical evidence that, even with proper choices on privacy settings, an individual is still unable to control how his/her own private information is disclosed in online communications, due to the behavior of others. Our findings also suggest that users do consider others’ privacy settings when peer-disclosing information about them, yet not enough support is provided in existing systems to help users identify potential peer-disclosures.

Second, unlike prior privacy studies which focused on theory-based positivist approaches and relied on self-reported data, this research pursues a data-driven method with observational data to test the real-world effectiveness of privacy choices. This helps address the

researcher-practitioner gap identified repeatedly in recent research, e.g., “some of us are increasingly raising doubts about whether we can relate what we have found in our research to what practitioners or policymakers truly experience in reality” [6].

The rest of the paper is organized as follows. The next section presents a review of the related literature. Following that, we define the research questions and describe our research methodology, including the data collection and analysis processes on Twitter. We then describe our research results and explain how these results address the research questions. We conclude the paper with discussions of implications, limitations, and future extensions of our work.

## 2. Literature review

### 2.1 Privacy self-management

The past decade has witnessed significant advances in understanding the human aspect of privacy research, from conceptualizing the norm of information disclosure [19] to studying the behavioral dynamics of privacy decision-makings (see [1] for a review). Given the complex contextual nature of privacy, many behavioral studies have repeatedly reported that users have difficulty making decisions for privacy self-management, due in part to their bounded rationality [22], which prevents them from systematically evaluating costs and benefits before making privacy decisions [1].

Also, in the past decade, much research has been done to design more effective mechanisms for privacy self-management. Such mechanisms include privacy nudges [3, 27, 32], which offer subtle yet persuasive cues to help users make the “right” decisions with minimum cognitive efforts; tools to facilitate the understanding of privacy policies [8, 24]; and better permission request schemes, like constructing permissions based on the purpose of use [28]. Recognizing that users’ privacy decisions often vary by demographics and context, there were also work [12, 13] that provide personalized support for privacy decisions based on predicted user preferences.

The research in this paper is orthogonal to prior research, as its focus is on the effectiveness of privacy choices on managing real-world privacy disclosure, *not* on the effectiveness of these choices on inducing a proper decision from an end user.

### 2.2 Collective privacy concerns on OSNs

Recognizing that privacy cannot be theorized solely from an individual perspective, researchers developed

the concept of collective privacy [10, 25, 31] to capture how one user’s decisions, e.g., the tagging of a friend at Facebook, may affect others’ privacy. Recent work on collaborative privacy [11] found the need for users to communicate privacy preferences with one another in order to solve conflicts and reach agreement on ownership, access, and extension of private information. Unfortunately, while existing studies have proposed behavioral approaches for users to ask each other for permission [14] or regulate boundaries with each other [29], these behavioral approaches function only at a small scale, and become impractical for OSNs like Twitter.

### 2.3 Re-identification risks

In today’s digital world, the rising threat to privacy is further increased by connecting information about one individual found in multiple data sources, both online and offline [7, 17, 23]. In the context of OSNs, researcher have repeatedly found evidence that today’s OSNs afford new capabilities to combine social identity elements with personal identity elements, which will improve the accuracy of the identity linkage techniques [16, 17, 20]. This increases the threats to individuals’ privacy and makes OSNs a gateway to access individuals’ personal information [2, 18]. Because of this linkage threat, no matter how mundane the information, its disclosure can have significant repercussions [21]. To capture the linkage threat, in this paper, we consider the typical identity attributes that can be used to link individuals’ online records with personal identities - name, gender, age, birthday, home address, etc.

## 3. Research questions

Recall from Section 2 recent research that chartered the conceptual framework for privacy peer-disclosure in OSNs [5, 11, 25] and verified its existence in real-world OSNs like Twitter [4]. What is missing in the literature, however, is a study of the interplay between peer-disclosure and a user’s privacy settings. As a result, we do not yet fully understand the effectiveness of OSN privacy settings in practice, as it is not clear whether the change of a user’s privacy setting would affect the behavior of others on peer-disclosing private information about the user. To address this gap, we study two research questions in the paper:

- RQ1. *After a user hides his/her OSN activities from public view, are there still new activities by the user’s connections on the OSN that reveal personal identity information about the user?*

- RQ2. Does a user's change to a more stringent privacy settings decrease the peer-disclosure of the user's private identity information in future activities of the user's connections (once these connections are made aware of the user's new privacy settings)?

To answer these questions, we need to first build a conceptual foundation that defines peer-disclosure in OSNs: what actions trigger peer-disclosures, under what privacy settings, and what private information may be revealed? We develop this framework in the next section before using it to enable our research design described in latter part of the paper.

#### 4. Conceptual framework: peer-disclosure

We develop the conceptual framework of peer-disclosure in OSNs by building a taxonomy along two orthogonal dimensions. One dimension captures the various OSN activities that trigger peer-disclosure, while the other dimension captures the different kinds of privacy leakage stemming from these activities.

##### 4.1 Peer-disclosure activities

Recent work [4] identified two main types of activities that lead to the peer-disclosure of user A's private information by user B: 1) the *proactive mentioning* of A in B's OSN activity, e.g., the tagging of A in a tweet or a photo posted by B; and 2) B's *reactive interactions* to A's activities, e.g., a reply posted by B to A's tweet, or B's comment on A's Facebook post.

Since the focus of this paper is on the effectiveness of A's privacy settings, we examined several popular OSNs to study how these settings affect the two types of activities, with the results summarized in Table 1. Each OSN examined has at least 800 million monthly active users as of June 2018.

**Table 1.** Summary of whether privacy settings of user A affect the 1) access privilege and 2) interface design of user B's 1) activities that mention A and 2) reactive interactions with A.

	Proactive Mentioning		Reactive Interaction	
	Access	Design	Access	Design
Twitter	No	Yes	No	Yes
Facebook	No	Yes	Possible	Possible
Wechat	No	No	Possible	Possible

Interestingly, none of the OSNs adjusts the access rules for B's proactive mentioning of A based on A's privacy settings. In other words, other users are free to

mention A no matter how A limits access to his/her own activities. Yet some OSNs adjust the design workflow - i.e., either A or B sees interface changes on B's mentioning of A - according to A's privacy settings. For example, Twitter displays a lock icon alerting B of A's privacy setting (see Figure 1). Facebook offers an option for A to be alerted when mentioned by another user, and lets A choose whether the mentioning appears on A's timeline. Note in the case of Facebook, if B has a public timeline, then the mentioning of A is publicly visible no matter which option A chooses. In other words, the access rules of B's activities remain the same.

For reactive interactions, whether the access rules change varies by OSN. Twitter makes no adjustment to access. If a public user B replies to A's protected tweet, then the reply will be public (just like the case when A is public), along with the fact that B is replying to A<sup>2</sup>. But even Twitter makes changes to the workflow design. For example, no user will be able to retweet (with or without comment) A's protected tweets, as the option is disabled on the interface when A switches to the protected mode.

For Facebook and Wechat, access to interactions is determined by privacy settings of the *initiator*. That is, if A is the initiator, then B's reactive interactions are governed by A's privacy settings. For example, B's comments on A's Facebook post follow the same access rules as A's original post. On the other hand, if A comments on B's post, then the initiator becomes B, and both B's post and A's comment follow the privacy settings of B. In this scenario, even when A sets default access to "Only Me", A's comments on B's posts are publicly accessible so long as B's post is public.

In summary, one can see from Table 1 peer-disclosure *might* happen in all three OSNs through being mentioned by others, interactions with others, or both. Mentioning is particularly dangerous because the user being mentioned may not be aware of it - losing privacy even after stopping all activities in an OSN.

##### 4.2 Leakage from peer-disclosure activities

Numerous types of private information may be revealed through peer-disclosure - some, like the direct reference of one's name, are obvious to anyone, while others, like vague references to common past experience, may only be recognizable by a selected few. Furthermore, what one individual considers as highly private information may look completely irrelevant to another [25]. Such complexity makes it impossible to capture all privacy peer-disclosure that may occur from OSN activities. As a result, it is

<sup>2</sup> A's original tweet being replied to remains hidden.

important for us to properly define the scope of peer-disclosure studied in this paper.

We focus on peer-disclosures that trigger a particular type of threat to privacy: the ability for an adversary to associate an OSN user with his/her real-world identity. That is, we focus on the disclosure of personal identity elements such as name, gender, age, birthday, home address, etc., which can be linked with external sources, e.g., credit reports, Whitepages.com, etc., to unveil the real-world identity of an OSN user.

Two types of leakages are common for these personal identity elements:

- *Explicit specification*, i.e., the direct exposure of one's identity element. An example is "visiting @A's home in Washington, DC", which explicitly states the city A lives in.
- *Implicit inference*, i.e., revealing an identity element without explicitly stating its value. For example, when B tags A in a "happy birthday" post, the birthday of A is revealed without being explicitly specified. Another example is when A is tagged in many tweets of B with geotags all in DC. Even though non-deterministic, one can infer with a high likelihood that A lives in the city.

Note the type of leakage is, in theory, orthogonal to the identity element being leaked. For example, besides implicit inference of birthday, explicit specification is also possible through a post: "@A born on Christmas 1979". In practice, however, the peer-disclosure of an element may be much more likely through one type vs the other - e.g., for birthday, implicit inference is more "natural" than explicit specification. Also note that either type of leakage may happen from not only the content of an OSN activity but also metadata. The aforementioned geotag (for inferring the city A lives in) and activity timestamp (for inferring A's birthday) are examples of the latter type.

## 5. Research methodology

We now describe the detailed methodology used in our empirical study of the two research questions. We chose Twitter as the OSN to study because its privacy settings mostly resemble a "clean" dichotomy - one either sets all OSN activities to be public (the default option) or hides (nearly) all from public view (the protected mode). This sharp contrast significantly simplifies the study of the research questions.

### 5.1 Data collection

**Finding protected users:** We started with a uniform random sample of 1 million Twitter users [15], and used the Twitter RESTful API, specifically the GET

users/show command, to find whether the user is in the protected mode. This process yielded a total of 4,715 Twitter users in the protected mode.

For the purpose of this paper, we also applied a subsequent filtering process. First, since the personal identity elements vary significantly from one country to another, we chose to concentrate on the US population, and therefore filtered out users outside the US. We did so by excluding a user if either location or time-zone attribute of the user profile indicates a location outside the country. Second, we also removed users who did not have meaningful activities on Twitter, i.e., those with either a tweet count of 0 or a total of 0 followers and followees. These statistics are shown by Twitter on the user profile page no matter if the user is in protected mode or not. After applying these filters, we were left with 2,608 protected users.

**Identifying past names:** To collect the peer-disclosure activities involving these 2,608 protected users, we faced an important technical challenge: Twitter allows users to change their user names. If we simply search historic tweets with a user's current name, we may miss many peer-disclosure activities that mention the user by past names. To ensure a proper study of peer-disclosure activities, we must first collect all past names used by a user.

Unfortunately, since the 2,608 users of interest are all in protected mode, we cannot retrieve their past tweets (which are hidden from public view) to unveil their past names. Thus, we resort to an alternative of finding *replies* to their tweets posted by other, public, users. Since Twitter API returns along with a public reply the numeric ID of the conversation originator (i.e., the protected user), which is a persistent identifier that remains the same no matter how the user name has been changed, we can associate a user name included in the content of a reply with the persistent user ID in the API-returned metadata, and thereby identify the past names used by a protected user.

Specifically, for each of the 2,608 protected users, we first collected all public tweets that mention their current name, identified all 100,632 public users who posted these tweets, retrieved all replies posted by these users, and then identified from these replies the old names used by the protected users. In total we identified 684 past usernames.

**Determining timeline of public-protected switch:** Another technical challenge we had to address before studying a peer-disclosure activity is that we did not know for certain whether the activity occurred after a user switched to the protected mode, or when the user's tweets were still public. Note that while Twitter reveals whether a given account is currently protected, it does not show, through either the web interface or the API, *when* the user started protecting his/her

tweets. As such, if we simply considered all peer-disclosure activities as evidence for privacy leakage under protected mode, we would be mischaracterizing some activities that indeed occurred when a user's tweets were still public. Furthermore, to study RQ2, we need to compare peer-disclosure activities when a user was public vs protected. As such, it is critical to establish the timing of the public-protected switch.

To address this challenge, we started by identifying upper and lower bounds for the timing of a user's switch from public to protected mode. Establishing an upper bound, i.e., when the user is definitely in protected mode, is straightforward: We verified all 2,608 users were in protected mode as of November 21, 2017 and continued monitoring their account status till June 1, 2018. As such, all accounts have an upper bound of November 21, 2017 unless they were found to later switch back to public mode.

It is much subtler to establish a lower bound, i.e., the most recent timestamp when a user was definitely public, because we cannot query a user's status back in history. As such, we had to leverage a covert channel created by a special design feature of Twitter: Starting in April 2014, any Twitter user can "retweet with comments" any public tweet, yet this option is disabled for any protected tweet. If the original tweet later became protected, the comments part of the retweet remains visible to the public<sup>3</sup>, along with a shortened URL pointing to the original (now hidden-from-public) tweet which, upon expansion, becomes

<https://twitter.com/user-name/status/tweet-id>

One can see that we can now infer from the URL the author of the original tweet. Thus, if we find that user A is the original author for a "retweet with comment" tweet of timestamp  $D_1$ , then we know for certain that A was *not* in protected mode at  $D_1$ , because otherwise the "retweet with comment" feature would have been disabled. This establishes a lower bound for the switch. Since the feature was introduced in April 2014, we would not be able to find a lower bound for users who switched prior to that date. In addition, a tacit assumption here is that the switch happened only once. For cases where a user switched from public to protected and then back, potentially multiple times, please refer to discussions in Section 7.

Finding the timeline of public-protected switch leads to further filtering of the collected users. For the upper bound, we filtered out 129 users who later switched back to the public mode. The remaining 2,245 users (and the peer-disclosures involving them) become our pool of study for RQ1. To answer RQ2, we need to further establish the lower bound, which we

---

<sup>3</sup> unlike a regular retweet, which would be hidden from public view when the original tweet becomes protected

were able to do for 198 out of the 2,245 users. These 198 users hence become our pool of study for RQ2.

## 5.2 Identification of peer-disclosure

For each protected user, we first collected all public tweets posted by its followers and followees, and then identified a subset of them that either mention or are part of an interaction with the protected user. This subset represents the potential peer-disclosure activities involving the protected user.

After manually examining a sample of the subset, we found six personal identity elements that have experienced peer-disclosures of significant frequencies: name, gender, location, photo, birthday, and age. For each type, we found evidence of peer-disclosure from both proactive mentioning and reactive interaction, and also both explicit specification and implicit inference.

We also found that the peer-disclosure of an element may occur at different levels of granularity, e.g., peer-disclosed names could be first name only, last name only, or full name; peer-disclosed locations could be precise GPS coordinates, a neighborhood, a city, or a state. Because of this varying granularity (and, as a result, the varying degree of privacy loss), we need a numeric measure of privacy leakage that works across different identity elements and different granularity levels, so as to fairly compare the degree of peer-disclosure for public and protected users.

To this end, we adopt the information-entropy-based measure in [4], which quantifies the amount of information (in *bits*) contained in the disclosed identity elements at a granularity level. For example, the disclosure of gender is  $\sim 1$  bit, while first name reveals more information - about 10.7 bits. While we refer readers to [4] for technical definitions of this measure, it is important to note that the measure captures the correlation between identity elements. For example, if both gender and first name are disclosed, the quantified amount is *not*  $10.7 + 1 = 11.7$  bits, but only  $\sim 10.9$  bits, because first name already reveals certain information about gender - hence the increase of only 0.2 bit for disclosing gender on top of first name. This ensures a fair comparison across the many combinations of identity elements peer-disclosed.

## 6. Research results: RQ1 and RQ2

### 6.1 RQ1

Table 2 depicts privacy loss from peer-disclosure for the 2,245 users in protected mode. Recall from Section 5.1 that we are certain these users were in

protected mode from November 21, 2017 to May 21, 2018 - hence this time period becomes what we study for RQ1. Table 2(a) shows the number of protected users who has an identity element peer-disclosed during this period through each of the four possible scenarios, while Table 2(b) future calculates the average amount of peer-disclosure (in bits) to capture the varying granularity of disclosures.

It is important to note that, in both Table 2(a) and (b), the numbers reported in the “overall” column or row are *not* the sum of the corresponding rows or columns because a user may have birthday disclosed through all four scenarios, yet all will be counted as one in the overall column. Also, as explained earlier in the paper, the entropy of (name, gender) is not equal to the sum of entropy for name and gender, because of the correlation between the two elements.

**Table 2.** (a) number of protected users with peer-disclosure; (b) entropy of peer-disclosure in bits. Each table consists of four types of peer-disclosures: Proactive mentioning, Explicit specification (P/E), Proactive mentioning, Implicit inference (P/I), Reactive interaction, Explicit specification (R/E), Reactive interaction, Implicit inference (R/I)

(a)	P/E	P/I	R/E	R/I	Overall
name	76	19	66	5	111
gender	70	124	98	125	220
location	9	60	0	31	74
photo	67	0	6	0	72
birthday	0	61	0	3	61
age	8	7	0	4	17
Overall	221	168	170	156	538

(b)	P/E	P/I	R/E	R/I	Average
name	10.7	12.96	10.7	10.7	11.14
gender	1	1	1	1	1
location	6	5	0	5	5.24
photo	9.3	0	9.3	0	9.3
birthday	0	9.02	0	8.5	9.02
age	6.3	6.3	0	6.3	6.3
Average	3	2.5	1.6	0.7	5.5

One can see from Table 2 that the answer to RQ1 is affirmative: Even after a user switches to the protected mode, thereby hiding his/her OSN activities from public view, it is still highly likely - in our study, 538 (24%) out of 2,245 protected users - for his/her connections’ *new* activities to reveal personal identity elements about the user. The amount of peer-disclosure is also significant - an average of 5.5 bits and median

of 11.2 bits. Specifically, 10% of these users have more than 24.3 bits peer-disclosed - sufficient to narrow the identity space from 325.7 million (the population of US) to just 15.77 individuals.

## 6.2 RQ2

To address RQ2, we compared the degree of peer-disclosure before and after a user switched to the protected mode. Since this study requires a lower bound on the time of a user’s switch to the protected mode, our study base was reduced to 198 users for whom we were able to find the bound. For each of these users, we focused on two time periods: 1) public period, which we set as the 6-month period immediately preceding our established lower bound, and 2) protected period, which we set as the 6-month period after November 21, 2017, the date we started monitoring the protected status of these accounts. For each user, we identified all peer-disclosure activities that happened during the two periods.

**Table 3.** Peer-disclosure (bits) before and after switch to protected mode: (a) Proactive mentioning before (P-b), after (P-a), and  $p$ -value for reduction (P- $p$ ), and the same for Reactive interaction (R-b, R-a, R- $p$ ); (b) the same for Explicit specification (E-b, E-a, E- $p$ ) and Implicit inference (I-b, I-a, I- $p$ ).

(a)	P-b	P-a	P- $p$	R-b	R-a	R- $p$
name	3.74	0.75	$10^{-6}$	4.53	2.70	.01
gender	0.32	0.09	$10^{-6}$	0.54	0.41	>.05
location	0.90	0.34	.05	0.41	0.76	>.05
photo	2.97	1.36	$10^{-3}$	0.08	0.08	>.05
birthday	2.27	0.73	$10^{-4}$	0.07	0.07	>.05
age	0.27	0.05	>.05	0	0.05	>.05
Overall	9.80	3.27	$10^{-8}$	5.47	3.90	>.05

(b)	E-b	E-a	E- $p$	I-b	I-a	I- $p$
name	6.14	2.40	$10^{-7}$	0.57	1.05	>.05
gender	0.47	0.24	$10^{-4}$	0.45	0.35	>.05
location	0.14	0	>.05	1.10	1.10	>.05
age	0.27	0.05	>.05	0	0.05	>.05
Overall	9.63	3.73	$10^{-10}$	4.37	3.32	>.05

In Table 3(a), we compared the two time periods for proactive mentioning vs reactive interaction. We also performed the paired-sample  $t$ -test for each combination of identity element and activity type to determine whether there is a significant statistical difference after the switch. For proactive mentioning, all elements except age exhibit a significant reduction

of peer-disclosure ( $p < .05$ ). Meanwhile, no identity element except name sees a significant reduction for reactive interaction. This sharp contrast shows that users react to the protected status of a friend differently when they proactively mention the friend vs reactively interact with the friend's tweets.

Table 3(b) depicts the comparison of the two time periods for explicit specification and implicit inference. Note that we exclude two identity elements, photo and birthday, because we found no implicit inference for the former and no explicit specification for the latter - rendering the comparison moot. Interestingly, a similar contrast emerges from the results: a significant reduction for explicit specification (overall  $p \sim 10^{-10}$ ) but not for implicit inference.

## 7. Discussions

### 7.1 Design implications

The results described in Section 6 have three main design implications. First, the results suggest that users do consider their friends' privacy preferences when engaging in peer-disclosure activities, as evidenced by the significant decrease of proactive mentioning or through explicit specification. The design implication of this finding is that an OSN should provide a user with information about the privacy preferences of other users involved in an OSN activity, in order to allow the user to consider such preferences.

Second, the results indicate that the decrease on peer-disclosure is much more pronounced on proactive mentioning than reactive interactions. This indicates the possibility that a user might not be aware of the potential violation of a friend's privacy when the user is merely responding to the friend's activities instead of initiating an activity on his/her own. The design implication here is that an OSN provider may alert users of such danger during interactions, e.g., by informing users that replies to a protected tweet will not be protected, but instead can be seen in public.

Finally, the results show little change on implicit inference regardless of the protected status of the user being mentioned, indicating the possibility that users might not be aware of such (tacit) disclosure of private information. Again, the design implication here is to raise awareness. Specifically, an OSN may proactively offer users a snapshot of their friends' public profiles, including whether a friend has ever revealed his/her birthday, location, etc., so users can make more informed decisions that affect others. For example, when a user posts a happy birthday message tagging a friend, the OSN may popup an alert dialog stating

"Your friend's birthday has never been disclosed here. Are you sure on posting this message?"

While our scope here is limited to examining the effectiveness of privacy self-management in OSNs, the findings may apply to other related areas as well. For example, when a user allows a photo-editing app to access photos taken with friends, the friend's private information becomes accessible to the app owner, incurring similar peer-disclosure as discussed in this paper. We leave the study of privacy self-management in these related domains to future work.

### 7.2 Limitations

**Approximated lower bound:** We acknowledge several limitations in our study. First, the way we established the *lower bound* on a user's switch from public to protected mode is approximate in nature, because we cannot completely rule out the possibility of repeated switches in the 6-month period preceding the lower bound. That is, the user might have switched from public to protected and the back to public before the date we identified.

A way to address the limitation is to simultaneously monitor many public Twitter users, hoping to capture the exact moment of their switch to protected mode. Unfortunately, the rarity of such switch events, coupled with the cost associated with the daily monitoring of numerous public accounts, makes this approach infeasible for our study. Fortunately, we observed from our monitoring of protected users that the repeated switch between public and protected is rare, as we only observed 129 switches over a 6-month period for the 2,608 users (4.9%). Thus, we believe this limitation does not affect the validity of conclusions in this paper.

**(In-)comprehensiveness of peer-disclosure:** We identified many ways for identity elements to be disclosed through implicit inferences, e.g., the date of a "happy birthday" tweet discloses birthday, while the semantics of "girls' night out" reveals the gender.

While our manual examination strove to enumerate all peer-disclosures through implicit inferences, it had two inherent limitations: first is that it hindered scaling up the study to more users, and second is the ad hoc nature of the examination which left the possibility of certain implicit inferences not being captured. For example, consider the tagging of a user in a tweet about movie "Twilight". One familiar with the movie could infer the tagged user's age range given the movie's adolescence target. But this may be missed by someone unfamiliar with the movie. Such uncertainty makes it infeasible to guarantee the identification of all implicit inferences.

Despite the lack of a comprehensiveness guarantee, it is our belief that the results of our study remain valid

even after considering other inferences because of two reasons: For RQ1, more inferences can only amplify the risk of peer-disclosure for protected users. For RQ2, we applied the same examiner and same standards when examining peer-disclosures before and after the switch. In future studies, we plan to address this limitation by first assembling a training dataset that consists of Twitter users with known identity elements and their tweets, then using a combination of Natural Language Processing (NLP) and machine learning techniques to auto-identify potential inference channels from the training data. Given the existing research demonstrating the correlation between personal identity elements and language characteristics like word choices [26], we conjecture this automated approach could improve the comprehensiveness of identified peer-disclosures with new inferences.

**Limited scale:** Our study is limited in scale by the query access limitation enforced by free Twitter APIs, i.e., 180 requests per every 15-minute window. This constraint limited the number of users we could track, especially given our need to search for all tweets mentioning every user we track, and further download everything posted by authors of these tweets. We plan to expand the scope of the study in future work to address this limitation.

**Different time windows for public and protected modes:** In studying RQ2, we compared the amount of peer-disclosure for a user before and after the user's switch from public to protected mode. Since the time window for activity measurement in public mode is bound to be *before* that for protected mode, doing so introduces potential extraneous factors explaining the change of peer-disclosure amounts, e.g., the overall decrease of peer-disclosure activities on Twitter over time. A between-subject design would address this limitation but requires a higher sample size than what our study currently has. Thus, we leave the study of this issue to a larger-scale study in future work.

**General limitations of using observational data:** We argued in earlier part of the paper that the usage of observational data helped us address the researcher-practitioner gap identified in privacy research [6]. Nonetheless, we also must admit that observational data also have limitations. For example, we cannot infer with certainty causation from observational data, due to reasons such as the existence of potential extraneous factors, e.g. increased privacy awareness due to news coverage during the observation period.

## 8. Conclusions

In this paper, we presented a data-driven study examining whether privacy self-management tools

provided by real-world OSNs are indeed effective for protecting a user's privacy. Specifically, we analyzed the public disclosure of a Twitter user's identity *after* the user chose to protect his/her account and eliminate public access to all tweets. Our findings show that even when in protected mode, users still have private information *continuously* disclosed on Twitter, mostly through the activities of the user's connections. This shows a key limitation of privacy self-management: its inability to control privacy peer-disclosure by others.

Our examination of how peer-disclosures change after a user switches from public to protected mode reveals two surprising patterns: 1) a sharp decline of proactive mentioning, but little change on reactive interaction; 2) a sharp decline on explicit specification, but little change on implicit inference. These findings show that other users *are* considering the switched user's privacy preferences when engaging in OSN activities, but they need further support on identifying peer-disclosures (in reactive interactions and through implicit inference). We discussed design implications of our findings for OSN providers. It is our hope that this work will inspire more data-driven studies on the state of OSN users' privacy in practice.

## 9. References

- [1] Acquisti, A., Brandimarte, L., and Loewenstein, G. (2015). Privacy and Human Behavior in the Age of Information. *Science*, Issue 6221, 509-514.
- [2] Acquisti, A. (2004). Privacy and security of personal information. In *Economics of Information Security* (pp. 179-186). Springer.
- [3] Acquisti, A. (2009). Nudging privacy: The behavioral economics of personal information. *IEEE security & privacy*, 7(6).
- [4] Alsarkal, Y., Zhang, N., & Xu, H. (2018). Your Privacy Is Your Friend's Privacy: Examining Interdependent Information Disclosure on Online Social Networks. In *Proceedings of the 51st Hawaii Intl Conf on System Sciences*.
- [5] Biczók, G., & Chia, P. H. (2013). Interdependent privacy: Let me share your data. In *Intl Conf on Financial Cryptography and Data Security*.
- [6] Belanger, F., & Xu, H. (2015). The role of information systems research in shaping the future of information privacy. *Information Systems Journal*, 25(6), 573-578.
- [7] Creese, S., Gibson-Robinson, T., Goldsmith, M., Hodges, D., Kim, D., Love, O., ... & Scholtz, J. (2013). Tools for understanding identity. In *IEEE Intl Conf on Tech for Homeland Security (HST)*.

- [8] Gluck, J., Schaub, F., Friedman, A., Habib, H., Sadeh, N., Cranor, L. F., & Agarwal, Y. (2016). How Short Is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In Symposium on Usable Privacy and Security (SOUPS).
- [9] Jia, H., and Xu, H. (2015). Measuring Individuals' Concerns over Collective Privacy on Social Networking Sites. Proceedings of ICIS.
- [10] Jia, H., and Xu, H. (2015). Examining Users' Collective Privacy Concerns on Social Networking Sites. Proceedings of ICIS.
- [11] Jia, H., & Xu, H. (2016). Autonomous and interdependent: Collaborative privacy management on social networking sites. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 4286-4297).
- [12] Knijnenburg, B. P., Kobsa, A., & Jin, H. (2013). Dimensionality of information disclosure behavior. *International Journal of Human Computer Studies*, 71(12), 1144-1162.
- [13] Knijnenburg, B. P. (2015). A user-tailored approach to privacy decision support. Univ of California, Irvine.
- [14] Lampinen, A., Lehtinen, V., Lehmuskallio, A., & Tamminen, S. (2011). We're in it together: interpersonal management of disclosure in social network services. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 3217-3226).
- [15] Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). The Tweets They Are a-Changin: Evolution of Twitter Users and Behavior. *ICWSM*, 30, 5-314.
- [16] Li, J., & Wang, A. (2011). Criminal identity resolution using social behavior and relationship attributes. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*.
- [17] Li, J., & Wang, A. (2015). A framework of identity resolution: evaluating identity attributes and matching algorithms. *Security Informatics*, 4(1), 6.
- [18] Madden, M., Fox, S., Smith, A., and Vitak, J. (2007). "Digital Foot-prints: Online Identity Management and Search in the Age of Transparency," PEW Research Center (<http://pewresearch.org/pubs/663/digital-footprints>)
- [19] Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev.*, 79, 119.
- [20] Park, S. H. and Huh, S. Y. (2012). "A Social Network-Based Inference Model For Validating Customer Profile Data," *MIS Quarterly*, 36(4), 1217-1237.
- [21] Stutzman, F., Gross, R., and Acquisti, A. 2013 Silent Listeners: The Evolution of Privacy and Disclosure on Facebook. *Journal of Privacy and Confidentiality: Vol. 4 : Iss. 2 , Article 2.*
- [22] Simon, H. A. (1982). Models of bounded rationality: Empirically grounded economic reason (Vol. 3). MIT press.
- [23] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05), 557-570.
- [24] Sathyendra, K. M., Wilson, S., Schaub, F., Zimmeck, S., & Sadeh, N. (2017). Identifying the Provision of Choices in Privacy Policy Text. In Proceedings of Conference on Empirical Methods in Natural Language Processing.
- [25] Shi, P., Xu, H., and Chen, Y. (2013). Using Contextual Integrity to Examine Interpersonal Information Boundary on Social Network Sites. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), pp.35-38.
- [26] Vogel, A., & Jurafsky, D. (2012, July). He said, she said: Gender in the ACL anthology. In Proceedings of the ACL-2012 Workshop on Rediscovering 50 Years of Discoveries (pp. 33-41).
- [27] Wang, N., Grossklags, J. and Xu, H. (2013). An Online Experiment of Social Applications' Privacy Authorization Dialogues, Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW).
- [28] Wang, H., Li, Y., Guo, Y., Agarwal, Y., & Hong, J. I. (2017). Understanding the Purpose of Permission Use in Mobile Apps. *ACM Transactions on Information Systems*, 35(4), 43.
- [29] Wisniewski, P., Lipford, H., & Wilson, D. (2012). Fighting for my space: Coping mechanisms for SNS boundary regulation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.
- [30] Xu, H., Teo, H. H., Tan, B.C.Y., and Agarwal, R. (2012). Effects of Individual Self-Protection, Industry Self-Regulation, and Government Regulation on Privacy Concerns: A Study of Location-Based Services, *Information Systems Research*, Vol. 23, No. 4, pp. 1342-1363.
- [31] Xu, H. (2012). Reframing Privacy 2.0 in Online Social Networks. *University of Pennsylvania Journal of Constitutional Law* (14:14), 1077-1102.
- [32] Zhang, B., & Xu, H. (2016). Privacy nudges for mobile Applications: Effects on the creepiness emotion and privacy attitudes. In proceedings of the 19th ACM conference on computer-supported cooperative work & social computing (CSCW).