Noname manuscript No.
(will be inserted by the editor)

Managing Longitudinal Exposure of Socially Shared Data on the Twitter Social Media

Mainack Mondal · Johnnatan Messias · Saptarshi Ghosh · Krishna P. Gummadi · Aniket Kate ·

Received: date / Accepted: date

Abstract On most online social media sites today, user-generated data remains accessible to allowed viewers unless and until the data owner changes her privacy preferences. In this paper, we present a large-scale measurement study focused on understanding how users control the longitudinal exposure of their publicly shared data on social media sites. Our study, using data from Twitter, finds that a significant fraction of users withdraw a surprisingly large percentage of old publicly shared data—more than 28% of six-year old public posts (tweets) on Twitter are not accessible today. The inaccessible tweets are either selectively deleted by users or withdrawn by users when they delete or make their accounts private. We also found a significant problem with the current exposure control mechanisms – even when a user deletes her tweets or her account, the current mechanisms leave traces of residual activity, i.e., tweets from *other* users sent as replies to those deleted tweets or accounts still remain accessible. We show that using this residual information one can recover significant information about the deleted tweets or even characteristics of the

Mainack Mondal, Johnnatan Messias
MPI-SWS, Germany
E-mail: {mainack, johnme}@mpi-sws.org

Saptarshi Ghosh
Department of Computer Science and Engineering, IIT Kharagpur, India
E-mail: saptarshi.ghosh@gmail.com

Krishna P. Gummadi
MPI-SWS, Germany
E-mail: gummadi@mpi-sws.org

Aniket Kate
Department of Computer Science, Purdue University, USA
E-mail: aniket@purdue.edu

This work is an extended version of the paper: Mondal et al., Forgetting in Social Media: Understanding and Controlling Longitudinal Exposure of Socially Shared Data, Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16), Denver, CO, USA, June 2016.

deleted accounts. To the best of our knowledge, we are the first to study the information leakage resulting from residual activities of deleted tweets and accounts. Finally, we propose two exposure control mechanisms that eliminates information leakage via residual activities. One of our mechanisms optimize for allowing meaningful social interactions with user posts and another mechanism aims to control longitudinal exposure via anonymization . We discuss the merits and drawbacks of our proposed mechanisms compared to existing mechanisms.

Keywords Longitudinal privacy · Exposure · Twitter · User behavior

1 Introduction

> *"every young person one day will be entitled automatically to change his or her name on reaching adulthood in order to disown youthful hijinks stored on their friends' social media sites". – Eric Schmidt [14]*

The unprecedented sharing of personal, user-generated content on online social media sites like Twitter and Facebook has spawned numerous privacy concerns for the users of the sites [5, 6, 10, 13, 16, 26]. In this paper, we focus on a dimension of user privacy that becomes more challenging to manage with the passage of time, namely, *longitudinal privacy*. Users' privacy preferences for sharing content are known to evolve over time [5, 6]. There can be many reasons for such temporal changes in privacy preferences – e.g., the sensitivity or relevance of shared content changes with time; the biographical status of users and their friend relationships change over time. The challenge of managing longitudinal privacy for a user refers to the difficulty in *controlling the exposure of the user's socially shared data over time*. This challenge becomes more complex over time as the set of contents shared in the past grows larger and new technologies like archival (timeline-based) searches make it easier to access historical content shared under outdated privacy preferences.

Two recent studies [5,6] surveyed social media users to check if these users are concerned about managing longitudinal exposure. They found that the users are indeed concerned about privacy of their old content. However these studies did not answer how the users are *actually* controlling longitudinal exposure of their content in real world. Hence, they provide a string motivation for us to investigate the longitudinal exposure control in real world.

Against this background, this paper asks and investigates the following *two* foundational questions related to *understanding* and *controlling* longitudinal exposure of user data in social media sites, respectively:

1. In practice, is there evidence for users changing their privacy preferences for content shared on social media sites 5 to 10 years in the past? If so, what is the extent of the change in longitudinal exposure of user data?
2. In practice, how effective are the mechanisms provided by social media sites to enable users to control the exposure of their shared data over

time? Could we improve the effectiveness of longitudinal exposure control mechanisms?

To address these questions, we have gathered extensive longitudinal data (over 6 years) from the Twitter social media site. Compared to the Facebook social networking site, the privacy preferences of users for messages (tweets) posted (tweeted) in Twitter are relatively simple. Specifically, Twitter provides three longitudinal exposure control mechanisms to their users—(i) users can withdraw their old tweets from public view by selectively deleting them (ii) they can withdraw all of their old tweets from public view by deleting their whole account and (iii) they can withdraw all of their old tweets from the public view by making their account private (old tweets only visible to followers of the private account). We elaborate each of these mechanisms in section 3.1. However, the simplicity of privacy choices in Twitter allows us to measure the temporal evolution of their users' privacy preferences by simply tracking the public visibility of users' tweets over time.

Our analysis of Twitter messages[1] reveals striking evidence of a significant fraction (∼35%) of all Twitter users changing their privacy preferences over time. Only a minority (∼8%) of all Twitter users selectively withdraw (i.e., delete or make them private) a small (∼10%) fraction of all their public posts. On the other hand, a sizeable fraction (∼27%) of all Twitter users withdraw all of their public posts older than a few (4-6) years. While a few recent studies have attempted to understand how user's privacy preferences might change with time through user surveys [5,6], to our knowledge, our work presents the first large-scale measurement study of how users actually change their privacy preferences in practice. Since our exploration is data driven (as opposed to user surveys), we could not investigate the user intentions behind the changes in privacy preferences. A limitation of our work lies in the assumption that these changes are driven by users' privacy concerns.

Our investigation of the effectiveness with which Twitter users control the public exposure of their tweets reveals a fundamental problem. Even after a user withdraws her public posts, the past interactions of her friends and other users with those posts (by the way of comments and replies) leave a trail of residual posts that remain on the site (as the residual posts are not authored by the same user, they cannot be withdrawn by her). We show that these residual activities are in many cases sufficient to recover significant amounts of information about the withdrawn posts, which is in serious conflict with the user's intent of withdrawing the post. These residual posts might reveal snippet of user's past lives (e.g., picture of wild partying where a deleted account is tagged) to the current acquaintances, including hiring managers. Thus residual activities pose a serious threat with the privacy of the users who withdraw their posts or their whole account and want the world to forget about their past actions. Our analysis of residual activities highlights this inherent flaw with the longitudinal exposure controls currently being pro-

[1] This study was conducted respecting the guidelines set by our institute's ethics board and with their explicit knowledge and permission.

vided to Twitter users. To make users more aware of the flaws in the existing exposure control mechanisms, we also design a Twitter app, deployed at `http://twitter-app.mpi-sws.org/footprint/`, where any one can login with their Twitter account and check the residual activities around their posts.

Having identified the limitations of existing longitudinal exposure controls, we discuss *why* devising a perfect solution to control longitudinal exposure is extremely difficult. Then we present an investigation into merits and drawbacks of a few advanced longitudinal exposure control mechanisms. Specifically, we focus on the recent trend towards ephemeral posts in new social media sites like Snapchat, where every post is timed to be deleted once it reaches a pre-set age (expiry time). The challenge with such ephemeral posts, however, lies in determining the "correct" pre-set deadlines for post deletion. We show that a different approach, where a post is deleted based on a pre-set duration of inactivity, offers users comparatively better control over their longitudinal privacy. Note that, the privacy breach by residual activities are taken care in our proposal (as well as in ephemeral posts), since we propose to delete every post (including the original and residual post) eventually.

2 Related work

In this section we explore the related work in this space along three axes.

Are users concerned about privacy of their old data? Understanding and improving privacy control in online social media sites garnered quite a bit of attention in recent times [5–8, 10, 11, 13, 16, 19, 20, 24, 26, 31]. The focus of these studies range from identifying regrettable / deletable content, to understanding the usage of privacy management mechanisms for sharing data, to designing better privacy management tools. However, there has been relatively little research on exploring the longitudinal privacy management mechanisms. Two recent studies [5, 6] surveyed tens to hundreds of users to explore how online social media users want to manage their longitudinal privacy for old content uploaded in the recent past (last week, month, year). The study in [5] performed a user survey and found that a user's willingness to share content drops as the content becomes old. Moreover, willingness of share further decreases with a life-change, e.g., graduating from college or moving to a new town. The other study [6] performed two surveys and discovered that users want some old posts to become more private over time and their desired exposure set for the content remained relatively constant over the years. Both of these studies indicate that users are, in general, concerned about the privacy of their old content, possibly because these content do not reflect who they are at present (possibly after a change in life). Hence, these studies provide a strong motivation for us to study at large scale how users in the real-world behave to address their privacy concerns.

How do users control longitudinal exposure of their old data? One natural way for a user to protect her longitudinal privacy is to delete her old

content. Some recent studies have focused on content deletion by users. For instance a PEW survey [18] on 802 teenagers found that 59% of respondents edited or deleted their content in OSNs. Almuhimedi *et al.* [4] reported the largest study so far on deleted tweets using real world data, however they only collected data which are deleted at most one week after posting. Specifically, they collected 67 million tweets from 292K users posted during a week, and found that 2.4% of those tweets are deleted within that week. Out of their set of deleted tweets, 89.1% were deleted on the same day on which they were posted. Moreover 17% of those deleted tweets were removed by the user due to typos or to rephrase the same tweet. However, note that, they primarily focused on content posted in the near past (no more than one week old) which were selectively deleted by the user. We will report later in this study how the exposure controls are quite different for the content posted in the near and far past, and show that the study [4] missed a large part of deleted tweets posted in far past (e.g., 6 years back).

A few other studies [12,17] explored the changing behavior of Twitter users over time. Out of them, Liu *et al.* [17] analyzed the collective tweeting behavior over time including deletion of content. They observed that social media users are either selectively deleting their tweets or deleting their entire account. However, they did not check if there are limitations of these mechanisms to control exposure. Neither did they explore the relative merits and drawbacks of different exposure control mechanisms. We explore these unanswered questions in detail.

What are some proposed mechanisms to help users control longitudinal exposure? Some recent studies mentioned possible mechanisms to improve the usability of longitudinal privacy mechanisms in OSNs. Bauer *et al.* [6] observed that users are possibly becoming more privacy-aware about their longitudinal data. This change in users' privacy concerns is further reflected by the advent and popularity of systems like Snapchat [2] which deletes all users' posts after a predefined expiry time. Aylan and Toch [5] proposed longitudinal privacy management mechanisms like allowing users to set expiration dates on content or having an archive feature for old content. We build upon these studies and propose a smart policy for content withdrawal, which dynamically tries to decide which content to delete or archive based on its longitudinal exposure.

Finally, a preliminary version of the present work has been published as a short paper [22]. The present work improves and extends the findings in [22]. For instance in section 5.1 we point out an additional problem with the existing longitudinal control mechanisms like age based withdrawal—they effectively remove history of any social activities and destroy archive of historical social media posts. Furthermore we propose a novel solution to improve the longitudinal exposure control in social media in section 5.3.1—content anonymization. Specifically, we propose a simple anonymization method (and two variations of the method) for controlling longitudinal exposure. We evaluate the effectiveness of our method using real world data in section 5.3.2 and show that

our proposal is robust against link based de anonymization attacks. Finally in section 5.3.3, we discuss the possible improvements of our anonymization mechanisms and point out concrete directions that future work can explore to improve the anonymization mechanism for controlling longitudinal exposure.

3 Understanding Longitudinal Exposure

In this section, we aim to understand how users are presently withdrawing their socially shared content to control longitudinal exposure. We start by answering the simple question – *what are the longitudinal exposure control mechanisms available today in Twitter, for withdrawing shared content?*

3.1 Exposure controls in Twitter

We found three distinct mechanisms of withdrawing socially shared content (tweets) in Twitter today:
1. Withdrawing tweets via selective deletion: The reasons for such deletion ranges from regrettable content in the tweets to simply correcting typographical errors or rephrasing [4].
2. Withdrawing tweets via deleting account: All tweets posted by a user can be withdrawn by deleting her whole account.
3. Withdrawing tweets via making account private: In Twitter, user-accounts are either 'public' or 'private'. Tweets posted by a public account are visible to anyone online, but tweets posted by a private account are visible to only the followers of that account, who must be approved by the private account owner before they can be a follower. Unlike Facebook[2], Twitter does *not* have sophisticated access control mechanisms whereby a tweet can be made visible to only a subset of one's followers. In Twitter, a tweet is either public to all users, or at least to all followers of the user who posted the tweet. Thus, if a user makes her account 'private', all tweets posted from this account are no longer available publicly.

Note that there is another factor that will result in tweets becoming inaccessible – if Twitter suspends a user's account for violating their terms of service, all tweets posted by that account will became inaccessible. However, we do not consider this factor as a mechanism for exposure control, since suspension is not carried out by the user herself.

To perform this study at scale, we needed to identify a large set of tweets that have been withdrawn by Twitter users. Additionally, we also needed to ascertain *why* a tweet has become inaccessible, so that we can ignore tweets that have become inaccessible due to Twitter suspending the users, and focus only on tweets that have been withdrawn by the users themselves. The rest of this section describes how we identified such tweets.

[2] Facebook's longitudinal exposure control mechanisms are more granular as observed by previous studies [16, 21]. Facebook users can choose to make their content available to only themselves, to their friends, subsets of friends, friends of friends or to general public.

Twitter error codes	Corresponding HTTP error codes	Twitter error message	Practical interpretation of Twitter error codes
179	403	Sorry, you are not authorized to see this status	User account made private
63	403	User has been suspended	User account suspended by Twitter
34	404	Sorry, that page does not exist.	Tweet (or user account) withdrawn
144	404	No status found with that ID	Tweet (or user account) withdrawn

Table 1 Error codes and error messages returned by the Twitter API when we try to access a tweet that has become inaccessible. The last column presents a practical interpretation of each error code.

Methodology for identifying tweets withdrawn by users: Our methodology consisted of taking a large set of tweets posted and archived in the past, and checking which ones have become inaccessible at the time of this study (October 2015). We observed that if we query the Twitter API with a tweet-id (a Twitter-generated unique identifier for a tweet) that was archived in the past when the tweet was public, if the tweet is inaccessible at present, the Twitter API sends back an error code and an error message as explanation. These error codes are customized by Twitter and are different from the normal HTTP error codes 404 (resource not found) and 403 (access forbidden) that are also obtained during this querying process. During our experiments consisting of querying for millions of tweet-ids (details given later), we noticed four distinct error codes that are shown in Table 1, along with the corresponding HTTP error codes, the corresponding error messages, and the practical interpretation of the error codes. These practical interpretations are based on the Twitter error messages and experiments performed using one of the author's Twitter account (as described below).

As shown in Table 1, the error messages accompanying codes 179 and 63 respectively identify the cases where the tweet has become inaccessible because the user made her account private, and where Twitter suspended the account. In this study, we will henceforth ignore the tweets that returned error code 63, since these tweets became inaccessible *not* due to user controlling their exposure, but rather due to Twitter suspending the users.

However, neither the Twitter official documentation[3] nor the error messages help to practically interpret the difference between the error codes 34 and 144. We experimented using the Twitter account of one of the authors of this paper, and observed that, both these error codes practically correspond to the case where the tweet has been withdrawn. However, these two error

[3] `https://dev.twitter.com/overview/api/response-codes`

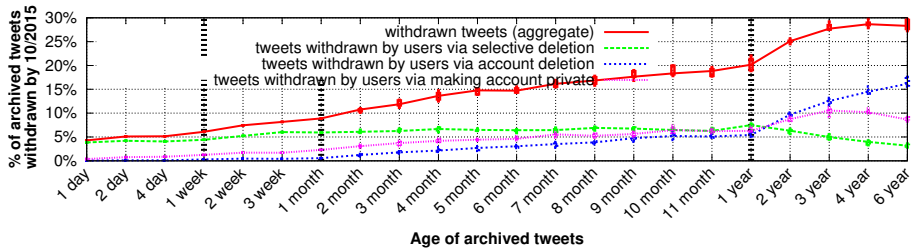

Fig. 1 Percentage of tweets in our sample of archived tweets that have been withdrawn as of October 2015. The age of a tweet is the difference between the time when the tweet was posted and the time of querying the Twitter API with the tweet-ids (October 2015). The amount of withdrawn tweets is increasing considerably over time – more than 28% of tweets posted 6 years back have been withdrawn today. The dotted vertical lines in the figure demarcate the points on the x-axis where the scale changes (days vs. months vs. years).

codes do *not* distinguish between the cases where the user selectively deleted a tweet and where the user deleted her account as a whole. To distinguish between these two scenarios, we further queried the Twitter API to check the *status of the user account* that had posted the tweet. The interpretation of codes is much simpler for user accounts (as compared to those for tweets) – the Twitter API returns HTTP code 200 OK for existing accounts, and error code 404 for deleted accounts.

Thus, by querying the Twitter API with archived tweet IDs (and the userids of users who posted the tweets), and observing the error codes returned, we can determine whether a previously public tweet has been withdrawn.

Limitations of our methodology: We do not know exactly *when* a tweet became inaccessible, i.e., how long after posting was it withdrawn. However, this limitation does not have much effect on the analyses we intend to conduct in the later sections. As we mentioned in the introduction, we also do not capture the user intention behind the withdrawal, i.e., we do not know exactly *why* a user withdrew her tweet or account (e.g., whether due to a change in life, due to change in friend relationships over time etc.). That said, we do view historical tweet withdrawal as being implicitly motivated by the desire for controlling longitudinal exposure of prior posts.

3.2 Longitudinal exposure of user data

To measure the longitudinal exposure of user data over the last *six years* from the time of the experiment (October 2015), we used two sets of archived data – (i) a near-complete crawl of Twitter done in September 2009 [9], consisting of 1.7 billion tweets posted by 54.9 million users, and (ii) a 10% random sample provided by Twitter (Gardenhose sample) collected from 2011 till the time of

this study. Note that all of these archived tweets were publicly shared when the data was originally collected.[4]

We fixed twenty-two time periods over the last six years, ranging from 1 day ago (from the date of our experiment in October 2015) to 6 years ago (see the x-axis in Figure 1). Then we randomly sampled 5,000 tweets from each of those time periods from our archived data.[5] We used the method described in the previous section on these tweet samples to check how many of the tweets from each time period have been withdrawn today due to exposure control of user data. We repeated the experiment over multiple consecutive days to make sure that the particular day examined was not an outlier (e.g., a holiday, the day a privacy news story broke, etc.). Specifically, for each of the time periods earlier than 2 months ago, we sampled 5,000 random tweets *per day* for a week around that time period and repeated our experiment.

1. How much of the archived data has been withdrawn? Figure 1 shows the variation in the percentage of tweets that have been withdrawn for each time-period. We show box and whiskers for time periods that are greater than or equal to 2 months, representing results from multiple days around those timestamps. The boxes show the span from first to third quartile and the whiskers signify minimum and maximum values. We observe that there is little variation (i.e, all the quartiles as well as minimum/maximum values are quite close to the median) among results from the repeated experiments over multiple consecutive days. Thus, our results did not vary much for tweets posted on samples picked from similar time in the past (e.g., from around 6 years back). Unless otherwise stated, we will report the median from the values obtained through the repeated experiments.

We discover that a substantial amount of past data has been withdrawn today. As shown by the solid red curve in Figure 1, the percentage of withdrawn tweets increases from 4.3% of the tweets archived 1 day ago to 28.3% of the tweets archived in 2009. Our observation suggests that users control the exposure for a significant amount of their past data. Hence the natural next question is: how do the different exposure control mechanisms account for this inaccessibility?

2. What is the relative usage of different control mechanisms for longitudinal exposure? Figure 1 further shows the variation of the percentage of tweets withdrawn via the three longitudinal exposure controls – (i) users selectively deleting tweets (green dashed curve), (ii) users deleting their account (blue curve), and (iii) users making their account private (pink curve). Surprisingly, we find that tweets posted from the near to far past have been withdrawn via very different exposure controls. Tweets posted in the near past

[4] We observed that Twitter provides a tweet in their random sample nearly instantaneously (within seconds) after a user posts the tweet. Consequently, there is at most a minimal chance that a user deleted a tweet even before it could appear in our random sample.

[5] We only considered original tweets (and not retweets) during sampling since our goal is to understand how much of the tweets originally posted by users are withdrawn today.

(e.g., 1 month ago) have mostly been withdrawn via users selectively deleting some of their tweets. However the percentage of tweets withdrawn via selective deletion quickly stabilizes over time (i.e., becomes more or less uniform for all time periods). On the other hand, the percentage of tweets withdrawn due to users deleting their accounts or making their accounts private, ramp up as we go further back in the past. In fact, these tweets account for the bulk of the older withdrawn tweets (e.g., 6 years back).

Specifically, out of 8.9% withdrawn tweets from September 2015 (1 month back), 5.9% consists of tweets selectively deleted by users and only 3% is contributed by users who deleted their account or made it private. Whereas, out of 28.3% withdrawn tweets posted in 2009, as much as 16.2% is contributed by users who deleted their account and only 3.2% by users who selectively deleted tweets.

It is important to note that prior studies on deleted tweets, e.g., by Almuhimedi *et al.* [4] *exclusively* focused on data from the near past (e.g., 1 week in the past), most of which are deleted shortly (within a few days) after they are posted. Hence, they ended up analyzing only the selectively deleted tweets, and missed the significant fraction of tweets posted in the far past that have been withdrawn due to users deleting their accounts or making the accounts private.

Summary: We analyzed the longitudinal exposure of socially shared data by measuring the percentage of tweets posted at different time periods in the past, that have been withdrawn as of today. We discovered that a surprisingly large fraction of old tweets has been withdrawn. Moreover, the exposure controls responsible for this withdrawal are very different for the near and far past. This global view motivates us to better understand privacy related behaviors at a user-level, i.e., *how are individual users controlling their longitudinal exposure?* We address this question next.

3.3 Understanding user behaviors

In this section, we assess individual users' behavior for controlling longitudinal exposure in the long-term. From the near-complete snapshot of Twitter data collected in September 2009 [9], we randomly selected 100,000 users who posted at least 100 tweets. For each selected user, we randomly sampled 100 tweets out of all the tweets posted by her (as obtained from the dataset). To simplify further analysis, we selected only the tweets that are in English, i.e., tweets in which at least 50% of the words appear in an English dictionary. Further, we ignored users who were later suspended, and the tweets posted by these users. We were left with 8,950,942 tweets (more than 89% of all tweets), posted by 97,998 users (97.9% of the users).

Using the methodology described earlier, we found that 29.1% of all the tweets that we checked have been withdrawn in the last six years, and these tweets were posted by 34.6% of our selected users.

3.3.1 Longitudinal privacy preferences of users

We start with categorizing our users into 3 distinct categories based on their usage of longitudinal exposure controls for withdrawing their tweets.

1. Non-withdrawers: users who did not withdraw any of their tweets. 65.4% of the users in our random sample fall in this class.

2. Partial withdrawers: users who only selectively withdrew some of their tweets. 8.3% of users in our sample are in this class. They have contributed 9.7% of the tweets that have been withdrawn.

3. Complete withdrawers: These are the users who have withdrawn all of their old tweets by either deleting their account or making their account private. As many as 26.3% of our selected users (25,751 in total) are in this class. Out of these users, 60.4% users have controlled exposure of their data by deleting their account, while 39.6% have made their account private. Out of all the withdrawn tweets in our sample, these users have contributed the bulk – 90.3% of all withdrawn tweets.

Table 2 shows the relative presence of each category of users in our dataset. We also show the breakdown of these users across different countries where only the top few countries (according to number of users) are shown.[6] The percentage of users with the different privacy preferences remains relatively constant across locations. This observation gives us some confidence that these privacy preferences are not location-specific, rather they are more universal.

Country	**Total users**	**Non withdrawer**	**Partial withdrawer**	**Complete withdrawer**
All	**97,998**	**65.4%**	**8.3%**	**26.3%**
US	43,412	65.4%	8.6%	26.0%
UK	4,870	69.7%	8.7%	21.6%
Brazil	4,576	60.8%	8.5%	30.7%
Canada	2,818	67.9%	10.7%	21.4%
Japan	1,740	73.2%	3.6%	23.2%
Australia	1,602	67.6%	7.9%	24.5%
Germany	1,439	67.7%	8.6%	23.7%

Table 2 A breakdown of all users by their privacy preferences as well as by their countries. Note that the breakdown of users by privacy preferences remains relatively consistent across countries.

One concern with our methodology is that, since we randomly sampled 100 tweets per user, we might potentially undercount the fraction of partial withdrawers. To check how serious this concern is, we repeated our experiments using *all* tweets posted by a set of users. However, due to the presence of some very active users, our sampled users posted more that 60 million tweets in total,

[6] We obtained the country of our users by leveraging location data of Twitter users gathered by Kulshrestha *et al.* [15]. They used the location and timezone field of the Twitter profile for inferring location of users.

category	Total #users	# users with inferred gender	% female users
Random population	97,998	65,438	50.3
Non-withdrawers	64,073	41,054	44.5
Partial withdrawers	8,174	5,667	55.7
Complete withdrawers	25,751	18,717	61.5

Table 3 Percentage of female users among different categories of Twitter users whose gender is inferred. The percentage of female users is higher among the partial and complete withdrawers than in a random Twitter population.

and given the rate limitations imposed by the Twitter API, it is very difficult to obtain the present status of all these tweets. Hence, we analyzed a slightly less active set of $\sim$ 97k random Twitter users from 2009, who posted between 10 to 100 tweets each. We repeated the same analysis as above considering *all* of their 2,622,808 English tweets. We found out that 13.6% of the users in this new random sample are partial withdrawers, which is only slightly higher than the fraction of partial withdrawers in our original sample of active Twitter users (8.3%).

We also found that, for a large majority of the users who posted between 10 to 100 tweets, the amount of information available is *not* sufficient for most of the analyses that we performed further (as described in the subsequent sections) due to lesser activity of these users. Hence, in the rest of our study, we will report results for our original set of 97,998 active users who posted 100 or more tweets each.

3.3.2 Correlating privacy preferences with demographics

Having identified users with different privacy preferences, we now check who these users are, by correlating the longitudinal privacy preferences of the users with their demographics. Twitter maintains only minimal demographic information for users, which includes only a profile bio and location. In spite of the absence of user-reported fine grained demographics information, there has been lot of prior work to infer different demographics characteristics for Twitter users [15, 23, 25]. We leverage this prior work to infer one important demographic for users from the available profile information – gender of these users. We focus on the gender since Tufekci *et al.* [29] noted a correlation between gender and privacy preferences of users in online social media.

We infer the gender from the self-reported first names specified in the user profiles using the methodology developed in [23]. Table 3 shows the percentage of female users among the users with different longitudinal privacy preferences. Interestingly, a majority of the partial and complete withdrawers are female, whereas the exact opposite is true for non-withdrawers. As a baseline, we checked that in a random sample of Twitter users, the percentage of males and females is similar. These results suggest that female users are controlling

exposure of their old data more than male users. This finding is also supported by an earlier study on Facebook [29] which reported that women are more likely than men to delete social media content.

Summary: We identify three distinct categories of users based on their individual use of longitudinal exposure control mechanisms. These privacy preferences of individual users do not vary significantly across countries. We also find that a majority of the content withdrawers are female.

After understanding the privacy preferences of different users, and observing the significant use of longitudinal exposure controls among them, we investigate our next question – are there any limitations of the current exposure controls?

4 Limitations of Existing Longitudinal Exposure Controls

Across online social media sites, the existing longitudinal exposure control mechanisms have an inherent limitation in the form of retained *residual activities* associated with a withdrawn post (e.g., a deleted tweet) or a withdrawn (deleted or private) account.

In these sites users frequently engage in conversations with other users, spurring interactions linked to their posts or to their accounts themselves (e.g., by mentioning a user in a tweet or by tagging a user in a Facebook post). Such interactions also include someone publicly replying to a specific post. When a user selectively deletes her post or withdraws her whole account, those old interactions (from others) associated with her withdrawn post or account become *residual activities* which still points to the withdrawn tweet or account. We show later in this section that, anyone today can collect a number of residual activities (e.g., *residual tweets* on Twitter) around both withdrawn tweets and accounts posted as far as six years back from the time of this study.

We acknowledge that such residual activities might exist even when a user deletes her recent post or withdraws her account created in recent past. However, intuitively, the amount of residual activities grows over time as an account stays longer in an online social media site, and consequently the associated privacy concerns become higher. Thus, we focus our analysis on the residual tweets around withdrawn tweets and accounts posted long back in the past (in 2009).

The presence of residual activities raises an immediate privacy concern – do the residual activities actually breach the longitudinal exposure control mechanisms? In other words, in the context of Twitter, can one recover information about selectively deleted tweets and deleted/protected accounts by simply collecting and analyzing the residual tweets associated with them?

4.1 Recovering information about selectively withdrawn tweets

We first focus on the *selectively* withdrawn tweets, which are deleted by their account holder while *retaining* some other tweets posted from their accounts. Specifically, we ask: what is the amount of the retained residual activities associated with these withdrawn tweets today, and what can we learn from them about withdrawn tweets?

4.1.1 Residual activities around withdrawn tweets

Data collection: We analyzed all the users who selectively withdrew one or more of their tweets from our random sample of 97,998 active users from 2009 (the same dataset as employed in Section 3.3). We then used Twitter search to collect conversations that mention any of those user accounts. Among these conversations, replies to a tweet still contain the tweet id of the tweet. Thus, we also identified the reply posts i.e., residual tweets involving those selectively withdrawn tweets from our dataset. Some examples of selectively deleted tweets with their corresponding residual tweets are shown in Table 4 (column 1 and 4).

Limitation of our data: Modified residual tweets like *RT@XYZ:<copiedPartialTweetText>* are easy to (programmatically) assign to withdrawn accounts (@XYZ) but not to particular withdrawn tweets. Therefore we included such residual tweets in the analysis of withdrawn accounts in Section 4.2, but not for the analysis of withdrawn tweets in this section. Thus, the data used in this section is effectively a lower bound on the residual activity around tweets. However, even so, we will show that one can still infer significant information about withdrawn tweets using this data.

How many residual tweets remain around the selectively withdrawn tweets?: In our dataset, a total of 8,174 users selectively withdrew their 253,853 tweets. We were able to collect 12,415 residual tweets posted in response to 9,738 of the withdrawn tweets. Although only 3.8% of all selectively withdrawn tweets have at least one residual tweet, these withdrawn tweets with residual activities were selectively withdrawn by a significant fraction of the users – 29.2% of 8,174 users who controlled longitudinal exposure by selective withdrawal. We further analyze the number of residual activities per withdrawn tweet. Figure 2 shows that, although a majority (89.2%) of these 9,738 selectively withdrawn tweets (with residual activities around them) have only one residual tweet, 3.8% of those tweets have more than two residual tweets. There is a maximum of 59 residual tweets around a single selectively withdrawn tweet in our data.

4.1.2 Recovering keywords from withdrawn tweets

We start by asking – can we recover meaningful words from the original withdrawn tweets just from the residual replies? To answer this question, we first

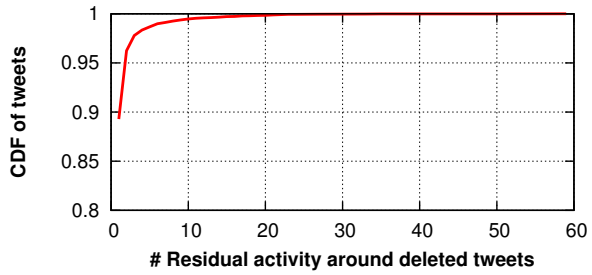

Fig. 2 Cumulative distribution function (CDF) for number of residual activities per selectively withdrawn tweet. Each of the withdrawn tweets have non-zero residual activity around it.

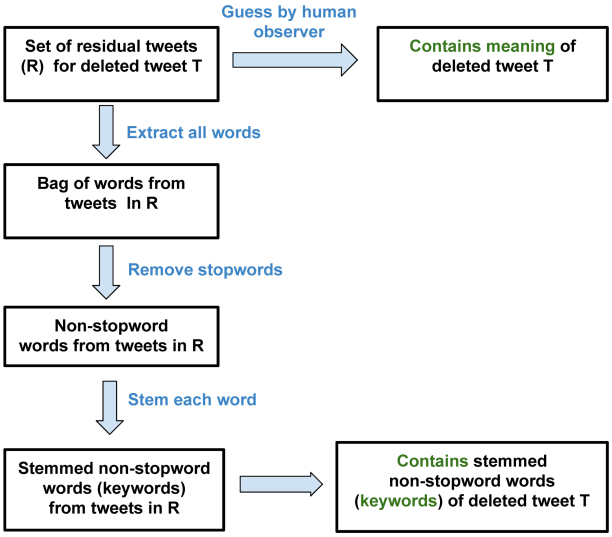

Fig. 3 Flowchart of our methodology for recovering keywords and meaning of deleted tweets from residual tweets.

removed all stopwords [7](no hashtags were removed in the process) from selectively withdrawn tweets and their associated residual activities, then stemmed the remaining words. We call the resulting set of words for a tweet *keywords*. We present a flowchart of the method of extracting keywords in Figure 3. We then checked what fraction of keywords from a withdrawn tweet also appears in the keywords from the set of residual tweets around it.

How many keywords can we recover from the withdrawn tweets?: Figure 4 shows the fraction of keywords shared by the withdrawn tweets and the residual tweets, as the number of residual tweets increases. We report the median values (unless otherwise stated) in this section, and the boxes in Figure 4 indicate the 25th and 75th percentiles. Note that we could recover

[7] We use a list of English stopwords and a list of Twitter-specific stopwords from [30].

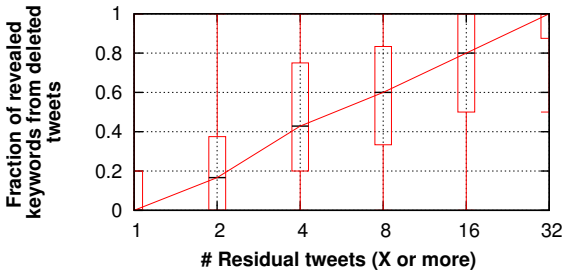

Fig. 4 Fraction of keywords that could be extracted for each of the withdrawn tweets (with at least one residual tweet) with varying number of residual tweets. The boxes indicate the 25th and 75th percentiles in the fraction, and the whiskers indicate the minimum and maximum values. The recovered keywords from withdrawn tweets increase with the number of residual tweets.

Original withdrawn tweet	**#Residual tweets**	**Example keywords from residual tweets**	**Example residual tweets**
Saw The Cove last night. Made me think about how much ALL animals need our respect – dolphins, cats, pigs, dogs, cows, chickens...	1	**cove**, **respect**, **animals**, extend, yeah, sea, recommending, veganfail, eat	"@[username] Yeah, but too bad "The Cove" doesn't extend that respect by recommending to not eat any animal from the sea"
[url] - Is it bad for you to eat unbaked cookie ? Hope not	3	**cookie**, **eat**, dough, batter, yummy, eveyone	"@[username] Cookie dough is awesome! Eat it up.", "@[username] i don't think so. isn't it like eating cookie dough? i do it with cake batter all the time. it's yummy"
What happened with Palin?	7	**palin**, resigining, alaska, safe, dearly, white, house, fantastic, definitely	"@[username] she's resigning. awww...", "@[username] she's going to act now....Nat'l Lampoon: Palin goes to Hollywood."

Table 4 Examples of withdrawn tweets, example keywords from the residual tweets, and actual examples of residual tweets. The keywords common in withdrawn tweets is shown in the bold font. As the number of residual tweets increases, their keywords give out more context about the withdrawn tweet.

16.7% of the keywords when the withdrawn tweets received two or more replies. Moreover, as expected, more residual tweets allow recovery of more information – the fraction of common keywords increases as the number of residual tweets increases.

Keywords revealed from the residual tweets: Table 4 shows some sample withdrawn tweets along with their residual tweets and the keywords gathered from the residual tweets. The keywords that also appear in the withdrawn tweets are highlighted using a bold font. Note that even if all the keywords

from residual tweets do not match the ones in the withdrawn tweet, they offer significant contextual information regarding the withdrawn tweet. This becomes more evident as the number of residual tweets increases. Specifically, in our examples residual tweets reveal the movie preference of the user (she saw"The cove"), her messy eating habit (eating raw cookie dough) or her political inclination (against Sarah Palin, vice president nominee of US). Note that the user originally wanted to withdraw all of this information by deleting the original tweet. This observation motivated us to consider another ambitious idea: *to what extent is it possible for a human observer to guess the meaning of a withdrawn tweet from the residual tweets?* Specifically, we asked human observers to guess a withdrawn tweet from its residual tweets, and then informally checked whether the meaning of the guessed tweets is qualitatively similar to the meaning of the original withdrawn tweet.

4.1.3 Recovering meaning of withdrawn tweets

Original withdrawn tweet	**#Residual tweets**	**Guessed tweet from AMT workers**		
		Guess 1	**Guess 2**	**Guess 3**
Saw The Cove last night. Made me think about how much ALL animals need our respect – dolphins, cats, pigs, dogs, cows, chickens...	1	The Cove has vowed to not eat any animals, good start!	Loved The Cove!	I think it's cool that the cove doesn't eat animal meat.
[url] - Is it bad for you to eat unbaked cookies? Hope not	3	Cook cookies? no thanks, I'll just eat them raw.	Are you sure I can eat this stuff? It's got raw food in it	I made cookie dough, but I can't seem to actually bake the cookies because I can't stop eating the dough!
What happened with Palin?	7	Sarah Palin finally stepping down, good day!	Read Sarah Palin's governorship resignation speech here: <link>	I wonder why Palin is resigning??

Table 5 Examples of selectively withdrawn tweets and the corresponding tweets guessed by AMT workers who were shown only the residual tweets for a withdrawn tweets. As the number of residual tweets increases, the AMT workers guessed the meaning of the original withdrawn tweet more closely.

Since guessing the meaning of a tweet automatically is a hard problem, we instead took help of human annotators from Amazon Mechanical Turk (AMT) for a preliminary demonstration. We used three AMT master workers from the USA for this survey. Each worker was first shown 5 example tweets and their replies. We first binned all of our selectively withdrawn tweets into five bins by the number of their residual tweets (i.e., tweets with $1, 2, \ldots, 5$ or more residual tweets) and selected ten withdrawn tweets from each bin. For our randomly sampled 50 withdrawn tweets, all the AMT workers were then shown the residual tweets of each withdrawn tweet and were simply asked to "Guess the original tweet". Finally we read through the guessed tweets and informally checked the (qualitative) resemblance between the meaning of the original withdrawn tweet and that of the guessed tweets. The simplicity of our AMT experiment demonstrates that a human observer can easily extract the meaning of deleted content by just reading the residual tweets. We present a flowchart of the method of guessing the meaning of deleted tweets in Figure 3.

Table 5 shows a part of the result from our AMT experiment.[8] As expected, when the number of residual tweets is small, the AMT workers were sometimes unsure about the meaning of the withdrawn tweet. Nevertheless, as the number of residual tweets increased, all the human observers guessed the meaning of the withdrawn tweets reasonably well (as reflected in their guessed tweets). This observation indicates that residual tweets often give out sufficient information for a human observer to guess the meaning of selectively withdrawn tweets.

Summary: We demonstrate that it is possible to recover both keywords and meaning from the withdrawn tweets by collecting and analyzing the available residual tweets associated with them. This is definitely a bad news for the users who wish to control exposure of their old post through selective withdrawal.

4.2 Recovering information about withdrawn accounts

Twitter users widely employ two mechanisms towards controlling longitudinal exposure of their accounts – some prefer to delete their accounts, while others prefer to make accounts private making their content inaccessible to a public observer. We collectively call these deleted or protected accounts *withdrawn accounts*. Here, we study two questions: what amount of residual activity around a withdrawn account is available, and what information does this residual activity reveal about the withdrawn accounts?

4.2.1 Residual activities around withdrawn accounts

We collected residual tweets around withdrawn accounts using a similar methodology as described in Section 4.1.1. We considered the withdrawn ac-

[8] For an interested reader to check the resemblance in meaning between the guessed and original tweets, we put our complete AMT evaluation result at `http://twitter-app.mpi-sws.org/soups2016/amt_guess.html`.

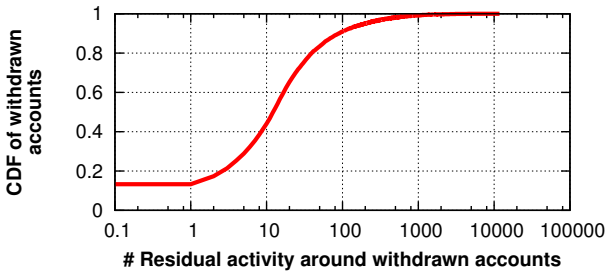

Fig. 5 CDF of number of residual activities per withdrawn account. More than 55% of withdrawn accounts have more than 10 residual tweets.

counts from our random sample of 97,998 users from 2009 (same dataset from section 3.3), and then used Twitter search to collect posts that mentions any of those user accounts. We limited our search to the period when the withdrawn accounts were active in our dataset, i.e., from the account creation date to the date of the last tweet appearing in our data.

How many residual activities remain around withdrawn accounts?: We collected a total of 1,403,716 residual tweets that mentioned 23,526 withdrawn accounts. In other words, a substantial fraction (91.4%) of the 25,751 withdrawn accounts have some residual tweets around them. We analyzed the number of residual activities around each account. Figure 5 shows that a significant amount of residual activities remain even at an individual account level – 55.9% of all withdrawn accounts have 10 or more residual tweets. Next, we ask what information can we recover about these withdrawn accounts, using both the residual tweets and the existing accounts that posted those residual tweets?

4.2.2 Recovering social connections

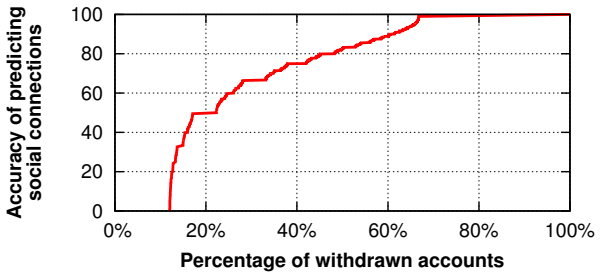

Fig. 6 The accuracy of our social connection inference with the percentage of withdrawn accounts for which we get this accuracy. For more than 30% of withdrawn accounts, all of their residual tweets came from their social connections.

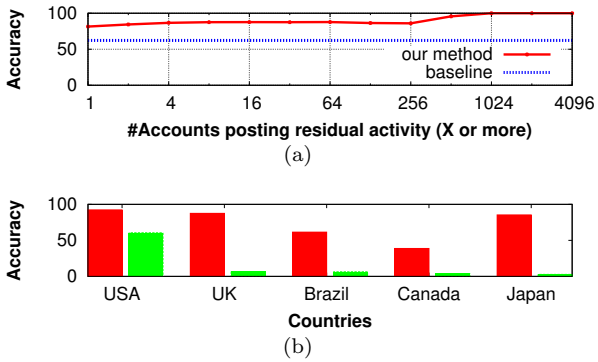

Fig. 7 7(a) Accuracy of our location inference leveraging residual activities. We can infer location with high accuracy and the inference is consistently better than baseline. 7(b) the accuracy for withdrawn accounts from different countries. First bar for each country is accuracy of our method and second bar is percentage chance that a random user will belong to that country.

We expect that two users converse mostly when they are socially connected. Thus, as a first test, we check if the users who mentioned a withdrawn account were connected to the withdrawn account by the follower-following relation. Cha *et al.* [9] had collected all the followers and followings of all Twitter users in 2009 and our withdrawn accounts are part of their dataset. Leveraging their collected data, we took all the social connections (both followers and followings) for each withdrawn account as our ground truth. Then we did a simple prediction: we predicted that each of the accounts mentioning a withdrawn account are either followers or followings of the withdrawn account. The accuracy of our inference for each user was: for what percentage of cases was our prediction correct?

Figure 6 shows the accuracy of our inference and for what percent of users we have a specific accuracy. Significantly, for 33.3% of the withdrawn accounts, the accuracy is 100%, i.e., all residual activities around these withdrawn accounts were posted by their social connections. For 48.3% of the withdrawn accounts, accuracy is more than 80%. Therefore, simply by checking who posted the residual tweets associated with a withdrawn account, we can recover some social connections for a significant number of withdrawn accounts.

A large number of existing studies pointed out that connected users in online platforms show homophily, i.e., have similar characteristics [3,28]. So we next check if we can recover some of the demographic attributes, like location, for the withdrawn accounts by leveraging the demographics of the accounts who contributed to the residual posts.

4.2.3 Recovering demographics

We here focus on whether we can infer the location of an withdrawn account from the location of the accounts who contribute to the residual activity

around the withdrawn account. As stated earlier, we obtained the ground truth country-level location for user-accounts from the study [15]. We then picked the most frequent location among the accounts which posted the residual tweets, as our predicted location for the corresponding withdrawn account. Our accuracy was decided by the number of withdrawn accounts for which our prediction was correct. As a baseline for comparison, we take the accuracy of a trivial predictor that selects USA as location every time (the most popular country in Twitter population).

Demographics prediction accuracy: Figure 7(a) shows the accuracy of our prediction with increasing number of user accounts associated with residual tweets. Significantly, when a withdrawn account has three or more accounts posting residual tweets around it, just by leveraging the residual activities we can infer the withdrawn account's location in 85.8% cases. This is consistently better than the baseline.

We also analyzed accuracy of our location inference for top five countries for the withdrawn accounts with some residual activities. The baseline accuracy for each country in this analysis was the accuracy of a predictor that outputs location based on the chance that a random Twitter user will belong to that country (computed using the full random sample of ∼98K users from Section 3.3). Figure 7(b) shows the comparison of accuracy for top five countries. We note that even for countries like Japan, where the chance of a random user coming from the country is as low as 2.25%, our inference is accurate for more than 87% withdrawn accounts.

4.2.4 Recovering topics of interest

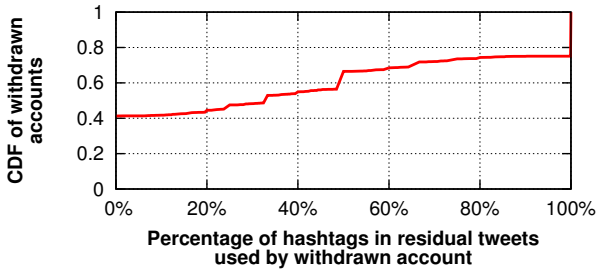

Fig. 8 The percentage of hashtags revealed by residual tweets that were originally also used by a withdrawn account. 25% of the withdrawn accounts, who ever used any hashtag in their tweets, used all of the hashtags revealed from their residual activities.

To recover potential topics the withdrawn accounts could have been interested in, we leveraged a special type of keyword – *hashtags*. Hashtags are words in tweets that starts with a '#' symbol and are included to provide the tweet a specific context. Practically hashtags are used to group together multiple

tweets on the same topic. For example, there were multiple tweets posted with "#iranelection" in 2009 to identify the topic of the tweet related to Iran election 2009.

Using data from [9], we determined that 3,855 accounts in our set of withdrawn accounts posted at least one tweet with a hashtag. Out of those, for 58.7% accounts (2,263 in total), the residual tweets revealed at least one of their hashtags, and in total 3,625 unique hashtags were revealed for these withdrawn accounts. This correlation encouraged us to further check what percentage of the hashtags revealed by the residual tweets were also used by the withdrawn accounts. Figure 8 shows our results: interestingly, in 25% of the cases, *all* the hashtags revealed by the residual tweets were also used by the withdrawn account.

User serial	**Topics**	**Hashtags used by withdrawn accounts, that are revealed by residual tweets**
1	Politics, Sports, Technology	#iranelection, #prisoners, #strike, #frenchopen, #tech
2	Politics	#conservativebabesarehot, #tea-party, #tcot, #obamacare
3	Sports, LGBTQ issues	#daviscup, #samesexsunday, #india, #lgbt, #followfriday
4	Sexuality, Entertainment	#furgasm, #nsfw, #gay, #shazam, #music
5	LGBTQ issues	#housing, #dcmetro, #protest, #gaymarriage
6	Politics	#immigrationreform, #iranelection, #peace #lgbt
7	Religion	#jesus, #truth, #idol
8	Sports	#grandrapids, #nascar
9	Sexuality	#hugeboner, #carchat
10	Sports, Entertainment	#collegefootball, #seinfeld

Table 6 Hashtags revealed by residual tweets for 10 withdrawn accounts. These users themselves used each of these hashtags. Also shown are some manually annotated topical categories these hashtags fall into. These hashtags give us an idea of what might be the topics of interest of the withdrawn accounts.

We further analyzed the hashtags revealed from residual tweets for some individual withdrawn accounts, and manually annotated the hashtag topics. Table 6 presents some example hashtags from the residual tweets of 10 users, who had used all of these hashtags in their now withdrawn tweets. As shown by our manual topical annotation of these hashtags, these hashtags shed light on the user's interests partially if not fully. Interestingly, some of these hashtags like "#iranelection", "#nsfw" might even be considered sensitive, while other hashtags such as "#daviscup", "#tech" or "#nascar" give away specific interests of the withdrawn accounts. This observation provides evidence that

the residual tweets still reveal information about what a withdrawn account was interested in, even when the account become inaccessible.

Twitter app to raise awareness about residual activities: To increase user awareness about their residual activities, we designed a Twitter app, using which any Twitter user can check what information about her account and individual tweets can be inferred by simply analyzing her residual activities on Twitter. We invite readers to use the app by visiting `http://twitter-app.mpi-sws.org/footprint/`.

Summary: We found significant evidence that the residual tweets and their associated user-accounts can be leveraged to at least partially recover the social connections, demographics (location) and even topical interests of the withdrawn accounts. Hence, the goal of the withdrawn tweet / account owners to control exposure of their (past) data cannot be achieved by the existing exposure control mechanisms. In the next section, we discuss the relative merits and demerits of a few exposure control mechanisms, and how such mechanisms can be improved.

5 Towards Better Longitudinal Exposure Control Mechanisms

Our analyses in the earlier sections show that a large number of users withdraw their past social content, but often a significant amount of residual information is left behind, which might lead to significant information leakage about withdrawn social content (and consequent privacy violation). This calls for an improvement of longitudinal exposure control mechanisms, which will directly increase the usability of such systems from a privacy perspective.

However, it must be understood that improving longitudinal exposure control mechanisms is a complex problem, as this has to take into consideration multiple (and sometimes contradictory) factors, such as the desire to retain some old content while allowing other content to be completely removed without a trace [6]. In fact, analyzing the effectiveness of such a mechanism might require a far richer understanding of many dimensions like incorrectly (not) limiting exposure of (non-)desirable content, potential privacy impact of such false flags, ownership of residual activities, ease of use and even user sentiment. Among these concerns ownership of content is a specifically tricky issue to address and varies from one social media site to another. In Twitter, the user who posts a tweet retains the right to alter the access control for them (e.g., via deletion). Consequently, the residual tweets are owned by the users who posted them and even if the original tweet is deleted the residual tweets might not be deleted (since the respective users did not delete them). In ephemeral social media sites like Snapchat [2], the system operator additionally deletes all the respective content (original as well as residual posts) after a predefined time. Hence, it is very unlikely that there is a silver bullet to solve all the problems with longitudinal exposure control. The longitudinal exposure control mechanisms that are being deployed in different online social sites today,

aim towards improving different dimensions of the problem, some of which we discuss below. We also propose two novel mechanisms for longitudinal exposure control, which addresses some of the limitations of the existing mechanisms.

5.1 Existing Mechanisms

1. Putting users in charge of controlling their longitudinal exposure: This mechanism is used in most of the popular online social media sites, including Twitter and Facebook, where the users are expected to control their own longitudinal exposure by withdrawing individual posts / accounts. On the positive side, this mechanism perfectly captures the user intent of retainment or withdrawal of specific content. However, as the previous section demonstrated, even when users withdraw their posts or accounts, the residual activity surrounding the withdrawn posts (authored by other users) could leak significant information about the withdrawn content.

It can be argued that withdrawing the residual activities along with the withdrawn posts and accounts is a natural solution to this issue or residual information. However, any such tampering of the content authored by other users (other than the one who specifically wishes to delete her content) raises several difficult questions associated with ownership and control of the content.[9]

2. Age based withdrawal: *Ephemeral social media sites* such as Snapchat [2] and Cyber Dust [1] offer a potential way out of the residual activity problem. On such sites, every message is associated with an expiry time after which the post is automatically withdrawn and becomes inaccessible to the users. Ayalon *et al.* [5] also suggested that the system operators of non-ephemeral social media sites can offer their users similar timed expiry option such that the posts will become inaccessible to the public after the expiry time.

Though this mechanism solves the problem of residual activities (since even all the residual activities will be inaccessible over time), it has two limitations.

1. First, the default expiry time used in such mechanisms is generally too small (e.g., few seconds or few minutes), which prevents any meaningful discussion around any post. Since the most interesting posts also get deleted after the expiry time, such mechanisms might not be preferred in sites like Twitter which promote social discussions. One might argue that this limitation can be overcome by simply setting the expiry time according to user's preference. Unfortunately, as noted in [6], users are generally poor at anticipating when a post should be deleted and thus there is little chance that this simple solution will work.
2. Second, age based withdrawal or any related longitudinal exposure control methods have a common shortcoming. They solve the problem of residual activities by deleting *all* posts, effectively removing history of any social

[9] For example, Twitter today automatically deletes re-tweets of a deleted tweet, but *not* replies or mentions generated by other users.

activities and destroying archive of historical social media posts. However there are a big problem with such removal of old posts—no archive of historical social media posts. Some social media sites enable researchers as well as social media analytics companies to study a huge volume of user generated data. A prime example of such OSM is Twitter. Twitter made a portion of user generated data (i.e. tweets) available to researchers and businesses via their streaming or search API [27]. Research studies use this Twitter data to solve problems ranging from understanding human behavior to detecting spam. Many of these studies use historical Twitter data, i.e., data posted recently or in the past (weeks, months or even earlier) in their research (e.g., while investagive user sentiment for last five US elections). Unfortunately, age based withdrawal or similar mechanisms makes any such research efforts impossible.

To that end, next we present two proposals to improve the longitudinal exposure control mechanisms.

5.2 Proposal 1: Inactivity-based withdrawal

Our proposal is based on a simple intuition—when a post becomes inactive, i.e., it does not generate any more interaction or receive any more exposure, the post can be safely withdrawn (deleted/archived/hidden) from the public domain.

Note that 'interaction' is a general term that can involve several tasks based on the social media site; e.g., it can mean sharing the post (e.g., retweeting in Twitter), replying to the post or even viewing the post by the original posting account or other users. Large social media operators today collect all of these interactions.[10] Hence, they can easily check if a post is inactive for more than T days (for any given definition of inactivity), and then the post can be withdrawn from the public domain. Also note that a user can be given various options for withdrawing her posts which become inactive; for instance, instead of fully deleting the posts, she may instead decide to limit access to the post to only select friends or may even anonymize the posts by removing any identifiable information. Here we generally consider withdrawal of posts from the public domain, and leave the details of the exact access control decisions to the social media operators.

Note that, alongwith improve usability, inactivity based deletion preserves the defense against residual activities that age based withdrawal offers. Eventually all the residual activities are also deleted (due to lack of any further interaction) over time. However, compared to age based withdrawal, our mechanism has the following advantages. First, the users need not be burdened with deciding expiry times of their posts. Second, this mechanism allows meaningful

[10] `https://support.twitter.com/articles/20171990#`
`https://www.facebook.com/help/437430672945092`

discussions around interesting posts, since the posts are withdrawn only after the discussion around them has died down.

Limitations of inactivity based withdrawal: However inactivity based withdrawal is a simple improvement over age based withdrawal. Specifically even in the case of inactivity based withdrawal, all the past posts are eventually deleted. Thus even in this case there is no preservation of part of historical content. Moreover this mechanism does *not* capture a user's intent to retain some old content even after it becomes inactive (e.g., because it had acquired large popularity, or because of some user-sentiment around a particular post). Another limitation of this mechanism is that, if a post is continuing to get interactions because it is controversial in nature, this mechanism would lead to the post remaining in the public domain. To address such issues, this mechanism should be coupled with other exposure control mechanisms such as a user being able to specifically withdraw some posts, or indicating her desire to retain a post even after it becomes inactive.

Even if a user wishes to adopt our proposed mechanism, a technical question needs to be addressed – how to select a value for T, the number of days after which a post will be withdrawn? With a very small value of T (say, 1 day), we may end up losing some valuable interactions; on the other hand, if T is too high (e.g., six years) users run a significant risk of someone digging up information about their past lives. Next, we demonstrate how the system operators can leverage the past interaction history to select an appropriate value of T.

Deciding an inactivity threshold: We ask a simple question in this direction: if we set a threshold of T days of inactivity before withdrawing a post, how much of the interaction generated by a post is likely to be lost? To that end we perform the following experiment. We randomly sample 700,000 tweets posted in the first week of November 2011, i.e., more than four years back. Note that all of these tweets are accessible today. In our experiment we take "retweets" as a proxy for generated interactions by a tweet. For a given tweet, we can obtain this interaction information directly from the Twitter API (unlike interactions like residual activities).

In our dataset, 30,014 tweets received at least one retweet and they received 74,705 retweets in total. We collect information about when each tweet received their retweets using the Twitter API, and simulate setting our inactivity threshold at T days, i.e. each of these tweets will become inaccessible after T days of not getting any retweets. We analyze the number of future retweets we would lose for different values of T.

Figure 9 shows that if we set our threshold to be too low, say 1 day, we will lose a significant 5.5% of all the retweets. However, if we set our threshold at only 180 days (i.e., decide that after six months of inactivity a tweet might be withdrawn from the public eye) then only 0.4% of the future retweets will be lost. Note that the parameter T need not to be global, and every user may choose her own value. In fact, the system operator can show a range of values of

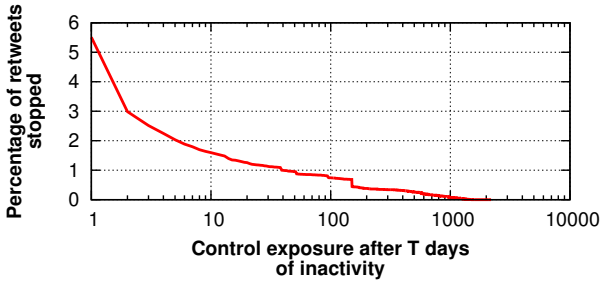

Fig. 9 Percentage of lost retweets if tweets were withdrawn after T days of inactivity, for different values of T. When T is set to 180 days only 0.4% of the future retweets will be lost.

Threshold in days	Inactivity based withdrawal		Age based withdrawal	
	#Retweets stopped	#Tweets these retweets came from	#Retweets stopped	#Tweets these retweets came from
1	4,117	1,584	7,798	1,681
7	1,342	556	2,678	587
30	842	317	947	339
90	609	235	744	243
180	300	181	579	193

Table 7 Comparison of age and inactivity-based threshold when both have the same threshold. Retweets of more active tweets are stopped by age-based threshold.

the threshold and point out the associated percent of stopped activities based on a user's past history, and allow the user can make an informed decision.

A comparison between the inactivity-based withdrawal and the age-based withdrawal: To demonstrate advantages of inactivity-based withdrawal over the age-based withdrawal, we also simulated age-based withdrawal policy with different thresholds over the same dataset of 700,000 random tweets and their retweets. Our age-based withdrawal policy is simple: after T days the tweet will be withdrawn and all future retweeting will be stopped. We closely investigate how many retweets will be affected by both these policies if we set same threshold. Table 7 shows the absolute number of retweets stopped and the number of tweets these retweets come from. It demonstrates that for the same threshold T, inactivity-based withdrawal stops comparatively fewer retweets than age-based withdrawal.

From our experiments, we make a more interesting observation: age-based withdrawal also affects tweets which generates lot of interaction (i.e., retweets) over a longer period of time, e.g, a tweet from the president of the United States. Let us take an example: Table 7 shows that when the threshold is set to

180 days, inactivity-based withdrawal stops 300 retweets from our dataset as it makes 181 tweets inaccessible. For the same threshold, age-based withdrawal makes 12 more tweets inaccessible (total 193), but stops 279 retweets from those additional 12 tweets, (i.e., on average 23 retweets per tweet). Notice that, by generating a lot of activity, popular tweets increase the usefulness of social content sharing systems. Thus, since age-based withdrawal might affect popular tweets, even with a high threshold it might not be suitable in the real-world adaptation. To demonstrate the effect of this issue, we measure actual time when a tweet will be withdrawn when we set an inactivity-based threshold of T days for different values of T.

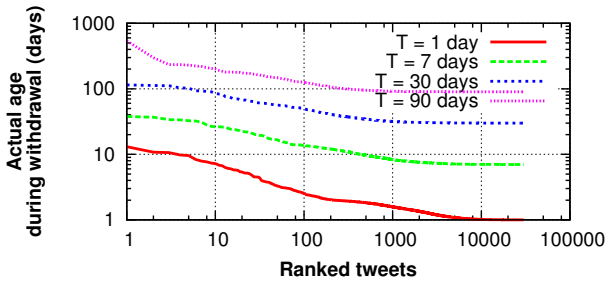

Fig. 10 Actual time when a tweet will be deleted when we set an inactivity-based threshold of T days.

In Figure 10, we plot the withdrawal age of the (inactivity-based) withdrawn tweets, and rank them in a sorted order based on their age. From the slope of these plots for different values of T, it is clear that the actual age of most tweets is significantly higher than their inactive age (or period).

Summary: We consider our inactivity-based withdrawal method to be an improvement over the age-based withdrawal, as it removes the need for a user to guess when her content should be withdrawn. Instead, the social site operator can present suggestions to users when a post becomes inactive, and facilitate the withdrawal. Next we propose anonymization as another effective method of longitudinal exposure control which preserves part of historical content.

5.3 Proposal 2: Anonymize withdrawn tweets

When a user withdraws her posts, she explicitly expresses desire for not to be associated with that content any more. Thus a simple trade-off between respecting her desire of withdrawal and still keeping part of the data would be an anonymizing scheme that anonymize and unlink a post from its publisher identity (the user who posted the withdrawn content).

In fact, there are multiple systems today that decouple publisher identities from content. Examples include fully anonymous social media systems

like Whisper (`http://whisper.sh/`) or YikYak (`https://www.yikyak.com/home`), which omits the concept of associating user identities with posts altogether. Moreover, another OSN operator, Reddit, already employ a version of anonymization for their withdrawn content (`https://www.reddit.com/wiki/privacypolicy`). Reddit simply removes the user identity in a withdrawn reddit comment and replace it by a "deleted" string.

Based on these observation, we propose the following idea: Social media sites should anonymize the withdrawn historical content to unlink the publisher identities from the tweets. This strategy provides a trade-off between user intent of withdrawal and keeping the archive of historical content—anonymization removes user identity from posts, detaching a publisher from her withdrawn content and existence of anonymized version of posts (as opposed to complete removal) preserved part of historical data archive. Again we will use Twitter as a platform to instantiate our proposal of anonymizing withdrawal social content (tweets in this case).

5.3.1 Anonymization scheme

We propose a simple anonymization scheme for tweets withdrawn by users; Twitter should just replace the publisher identities (e.g., publisher id, publisher's user-name) in the withdrawn tweet with a random string. In this way the studies that leverage the tweet content can still analyze the tweet text including urls and hashtags. Note that, here we consider tweets withdrawn by their publishers. Twitter can choose to keep the tweets they themselves withdraw (e.g., by suspending accounts) as it is (perhaps with a flag that this tweet is suspended). After the withdrawn historical tweets are anonymized, a researcher can still collect historical tweet data, and can obtain the urls, hashtags and words to continue her analysis. So the impact on the data quality for content analysis will decrease.

Limitations of our scheme: We acknowledge that our proposal is not a silver bullet. Specifically it does not address three issues: First, since all user identities are detached from their tweets, researchers focusing on specific parts from the population might not be able to collect data from those parts (e.g., collecting data from all female users posting about Brexit). However we believe that, this is a trade-off that a researcher have to make while respecting user intent of data withdrawal. Second, personally identifiable information (PII) might still remain in the tweets (e.g. as proper nouns or even writing styles) and our simple anonymization scheme will not remove them. Twitter might provide users the option to remove this PII by automatically identifying them. However finding the PII is highly context dependent and we leave it to future work for improving this aspect. Finally, any anonymization scheme still results in partial loss of information from social media sites and might create problems in some cases. For example, if some researcher wants to study network properties, then our anonymization scheme might be problematic for him, since

our anonymization scheme omits the network structure of users contributing to residual activities of deleted tweets.

5.3.2 Likelihood of de-anonymizing a publisher using network structure

There is another concern that our scheme needs to address. We observed that, Twitter is a social network and people converse about published content (e.g., in the form of replies to tweets). However while anonymizing a withdrawn tweet, Twitter can not simply anonymize these conversations too; that will raise a complicated ethical concern — these conversations are posted by users other than publisher of the withdrawn tweet, so, ideally, anonymizing the conversations requires all the conversing user's explicit consent. Still, these conversing users are highly likely to be connected to the original publisher in Twitter (by follower/following relations) and might reveal publisher identity thorough the network structure of the social graph. So an obvious question is *How likely* is it for an analyst to identify publisher of an anonymized tweet by simply looking at the social connections of the users conversing (e.g., replying) with the anonymized tweet. Next, we will thoroughly investigate this question using real world data.

We will first describe our dataset of Twitter conversations around withdrawn content. Then we will consider two possible implementations of our anonymization scheme: (i) *anonymization per withdrawn tweet* – each withdrawn tweet is anonymized independently, i.e., publisher identities in each withdrawn tweet is replaced by a unique random string and (ii) *anonymization per publisher* – where *all* withdrawn tweets from the same publisher, publisher identity is replaced by same random string.

Note that anonymization per publisher, although preserves more data quality (since all withdrawn tweets from same publisher and conversations around them can be grouped and analyzed in studies), it also leaks more information about user identity (though information about multiple people conversing with tweets from same publisher). We will investigate the likehood of deanonymizing publisher using network structure in both of these cases.

Dataset to evaluate anonymization scheme: For our analysis in this part we require a dataset of withdrawn tweets from a large sample of Twitter users and the conversations around these tweets. We leverage the same dataset mentioned in section 3.3 and section 4.1.1. Recall that our dataset contains 8,950,942 historical tweets, posted by 97,998 users. Furthermore 33,925 (34.6%) users withdrawn 2,605,317 (29.1%) tweets from this collection. Aside from historical tweets, there are 41,618 conversations (tweets posted as replies) around 36,796 of these withdrawn tweets from 7,964 publishers (23.5% of the publishers who withdrew their tweets). Our data also contains the social connections of these conversing users as well as the tweet publishers.

Using this dataset we seek answer to the question: How likely is it that the original publisher of a withdrawn tweet is revealed by simply looking at the social connections of the conversing users? Or in other words, how likely is it

that the original publisher of a withdrawn tweet is the *only* common neighbor of the conversing users in the Twitter social graph?

Deanonymizing a publisher when anonymization is done per withdrawn tweet: Recall that when each withdrawn tweet is independently anonymized, the user identity in each withdrawn tweet is replaced by a unique random string. So an analyst can only identify that conversations around each withdrawn tweet are addressed to a particular publisher. In that scenario, we take each withdrawn tweet and collect the social connections (both followers and followings) of the users who conversed with that tweet. Then for each withdrawn tweet we check the number of common social connections for the conversing users.

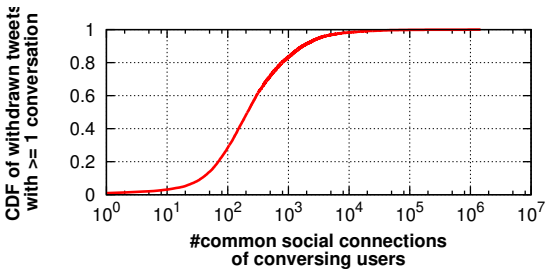

Fig. 11 The CDF of withdrawn tweets with one or more conversations. Only 0.9% of the tweets have one common neighbor within conversing users. Also 98.6% of the 2,605,317 withdrawn tweets did not spur any conversation (not included in the Figure).

The result is shown in Figure 11. We note that only for 0.9% of the withdrawn tweets with one or more conversation, there is exactly one common connection between the conversing users, and 96.9% of withdrawn tweets have more than 10. We concentrate on the withdrawn tweets, where there is only one common connection between the conversing users. We found that, only for 319 withdrawn tweets, the identities of 168 original publishers are revealed, i.e., those publishers are the only common social connection of the conversing users around those tweets. In other words, 99.5% of the 33,925 tweet publishers who withdraw their tweet can not be de-anonymized using the social connections of conversing users if Twitter leverage a simple *per withdrawn tweet* anonymization scheme.

We investigate further and find the main reason for this low likelihood of deanonymization: 98.6% of the 2,605,317 withdrawn tweets did not spur any conversation around them and 1.3% received only 1 conversation. Thus, for most of the publishers, an analyst does not have enough information from the social connections of the conversing users for revealing publisher identity.

Now we investigate the likelihood of de-anonymization if we replace user identities of all withdrawn tweets belonging to a particular publisher with a

random string, i.e., the random string is not unique for each withdrawn tweet, but for each publisher.

Deanonymizing a publisher when anonymization is done per publisher: In this case an analyst can identify and group multiple withdrawn anonymized tweets belonging to a particular anonymized user id. Moreover she can also identify that *all* the conversations around those withdrawn tweets are addressed towards a particular user. Thus, intuitively, an analyst have more information available for deanonymizing a publisher. We implemented this scheme and check, in how many cases, the publisher is the common social connection of all the conversing users around all of the publisher's withdrawn tweets.

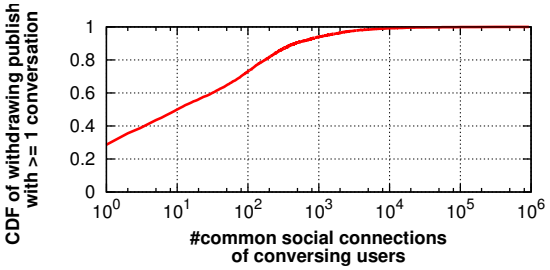

Fig. 12 The CDF of withdrawing publishers with one or more conversations. 23.5% of these publishers have exactly one common connection between conversing users. Note that, 76.5% of the 33,925 publishers with withdrawn tweets did not have any conversation (not included in the Figure).

We found that, 23.5% out of 33,925 publishers who withdrew tweets, have one or more conversation in total around her withdrawn tweets. Figure 12 shows the number of common connections between the conversing users for the publishers with one or more conversation around their withdrawn tweets. We note that for 28.7% of these publishers, there is one common connection within the conversing users, and for 50% publishers the number of common connections is more than 10.

We again focus on these publishers for whom there is exactly one common connection between the conversing users around her withdrawn anonymized tweets. We found that 1,784 publishers, i.e, 5.3% of the 33,925 publishers who withdrew their tweets, can have their identity de-anonymized using social connections of the conversing users. This fraction is certainly higher than the case when each withdrawn tweet is independently anonymized, but this anonymization scheme still protect identities of 94.7% of the publishers who withdrew their tweets.

5.3.3 Improving our anonymization scheme: Future directions

So far, we considered a possible anonymization strategy for Twitter as a longitudinal exposure control mechanisms which partially retains the quality of historical data. Our strategy involved simply anonymizing the withdrawn tweets by replacing user identities with random string. We found that, this strategy is robust against deanonymization by leveraging Twitter social graph and the users conversing with particular tweets (via replying to tweets). In fact in case of per withdrawn tweet anonymization 99.5% of publishers and in case of per publisher anonymization 94.7% of the publishers remains anonymized against such deanonymization attacks. The social media operators can choose which anonmization scheme will be suitable for their service, based on how much data quality they want to preserve. They might further improve their anonymization schemes, e.g., by removing personally identifying information (PII) while anonymizing withdrawn posts. We leave exploration of these further schemes as a potential direction of future work. Finally, an intermediate step before deploying this mechanism in platform like Twitter would be to give them the choice of both "anonymize" and "delete" old posts. This step would be useful in measuring user response regarding anonymization.

6 Conclusion

In this paper, we explored a dimension of user privacy that becomes more challenging to manage with passing time, namely, longitudinal privacy. Specifically, using extensive data from the Twitter social media site, we studied whether online users employ longitudinal exposure control mechanisms in real world to limit exposure of their old data. We find that a surprisingly large fraction (28%) of tweets posted in the far past are withdrawn by users today. After exploring the usage of existing privacy mechanisms by individual users, we find a significant problem with mechanisms to control data exposure today – social media sites retain residual activities around withdrawn content, which can be used to recover various important information ranging from social connections to user interests and even parts of the withdrawn content. We proposed two improved exposure control mechanisms—First, *inactivity based withdrawal* – an embodiment of the simple idea that old content can be safely withdrawn when it does not generate any more activity. We show its benefits for controlling longitudinal exposure over existing age-based exposure controls. Second, anonymizing historical withdrawal content, which improves the data quality of historical data by preserving parts of withdrawn content.

Our two proposals aim to optimize for two different dimensions of longitudinal exposure control—privacy and availability of data. Inactivity (and age) based withdrawal optimizes for privacy (since all data will eventually be deleted), whereas anonymization optimizes for availability. We acknowledge that comparing these two approaches (and subsequently choose one method over another) is difficult and requires throughly understanding (personalized)

user preferences in different contexts. In fact a concrete future research goal is to evaluate these approaches in the context of different social media sites and different types of content. We stress that, our proposals are not silver bullets to provide a "one size fits all" solution to the problem of improving longitudinal exposure control mechanisms. In fact, our study demonstrate the need and the scope of further research in this space. Specifically we identify a broad future research venue—researchers should undertake more detailed empirical data driven studies (spanning multiple social media sites) to design improved longitudinal exposure control mechanisms for socially shared data.

References

1. Cyber Dust. https://www.cyberdust.com/ (2016)
2. Snap chat. https://www.snapchat.com/ (2016)
3. Aiello, L.M., Barrat, A., Schifanella, R., Cattuto, C., Benjamin, M., Menczer, F.: Friendship prediction and homophily in social media. ACM Transactions on the Web **6**(2), 1559–1131 (2012)
4. Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., Acquisti, A.: Tweets are forever: A large-scale quantitative analysis of deleted tweets. In: Proceedings of the 16th Conference on Computer Supported Cooperative Work (CSCW'13) (2013)
5. Ayalon, O., Toch, E.: Retrospective privacy: Managing longitudinal privacy in online social networks. In: Proceedings of the 9th Symposium on Usable Privacy and Security (SOUPS '13) (2013)
6. Bauer, L., Cranor, L.F., Komanduri, S., Mazurek, M.L., Reiter, M.K., Sleeper, M., Ur, B.: The post anachronism: The temporal dimension of facebook privacy. In: Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society (WPES'13) (2013)
7. Bernstein, M.S., Bakshy, E., Burke, M., Karrer, B.: Quantifying the invisible audience in social networks. In: Proceedings of the 31st SIGCHI Conference on Human Factors in Computing Systems (CHI'13) (2013)
8. Besmer, A., Lipford, H.R.: Moving beyond untagging: Photo privacy in a tagged world. In: Proceedings of the 28th SIGCHI Conference on Human Factors in Computing Systems (CHI'10) (2010)
9. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: the million follower fallacy. In: Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM'10) (2010)
10. Dey, R., Jelveh, Z., Ross, K.W.: Facebook users have become much more private: A large-scale study. In: Proceedings of the 10th Annual IEEE International Conference on Pervasive Computing and Communications (perCom'12) (2012)
11. Hoadley, C.M., Xu, H., Lee, J.J., Rosson, M.B.: Privacy as information access and illusory control: The case of the facebook news feed privacy outcry. Electronic Commerce Research and Applications **9**(1), 50–60 (2010)
12. Jain, P., Kumaraguru, P.: On the dynamics of username changing behavior on twitter. In: Proceedings of the 3rd IKDD Conference on Data Science (CODS'16) (2016)
13. Johnson, M., Egelman, S., Bellovin, S.M.: Facebook and privacy: It's complicated. In: Proceedings of the 8th Symposium on Usable Privacy and Security (SOUPS'12) (2012)
14. Jr., H.W.J.: Google and the search for the future. http://www.wsj.com/articles/SB10001424052748704901104575423294099527212 (2010)
15. Kulshrestha, J., Kooti, F., Nikravesh, A., Gummadi, K.P.: Geographic dissection of the twitter network. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12) (2012)
16. Liu, Y., Gummadi, K.P., Krishnamurthy, B., Mislove, A.: Analyzing facebook privacy settings: User expectations vs. reality. In: Proceedings of the 11th ACM/USENIX Internet Measurement Conference (IMC'11) (2011)

17. Liu, Y., Kliman-Silver, C., Mislove, A.: The tweets they are a-changin': Evolution of twitter users and behavior. In: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14) (2014)
18. Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., Beaton, M.: Teens, social media, and privacy. http://www.pewinternet.org/2013/05/21/teens-social-media-and-privacy/
19. Madejski, M., Johnson, M., Bellovin, S.M.: The Failure of Online Social Network Privacy Settings. Tech. Rep. CUCS-010-11, Department of Computer Science, Columbia University (2011)
20. Mazzia, A., LeFevre, K., Adar, E.: The pviz comprehension tool for social network privacy settings. In: Proceedings of the 8th Symposium on Usable Privacy and Security (SOUPS'12) (2012)
21. Mondal, M., Liu, Y., Viswanath, B., Gummadi, K.P., Mislove, A.: Understanding and Specifying Social Access Control Lists. In: Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS'14) (2014)
22. Mondal, M., Messias, J., Ghosh, S., Gummadi, K.P., Kate, A.: Forgetting in social media: Understanding and controlling longitudinal exposure of socially shared data. In: Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16) (2016)
23. Mullen, L.: Predicting gender using historical data. https://cran.r-project.org/web/packages/gender/vignettes/predicting-gender.html (2015)
24. Petrovič, S., Osborne, M., Lavrenko, V.: I wish I didn't say that! analyzing and predicting deleted messages in twitter. CoRR **abs/1305.3107** (2013)
25. Sloan, L., Morgan, J., Burnap, P., Williams, M.: Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PLoS ONE **10**(3), e0115,545 (2015)
26. Stutzman, F., Gross, R., Acquisti, A.: Silent listeners: The evolution of privacy and disclosure on facebook. Journal of Privacy and Confidentiality **4**(2) (2012)
27. team, T.: The streaming apis. https://dev.twitter.com/streaming/overview
28. Thelwall, M.: Homophily in myspace. Journal of the American Society for Information Science and Technology **60**(2), 219–231 (2009)
29. Tufekci, Z.: Facebook, youth and privacy in networked publics. In: Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM'12) (2012)
30. Zafar, M.B., Bhattacharya, P., Ganguly, N., Gummadi, K.P., Ghosh, S.: Sampling content from online social networks: Comparing random vs. expert sampling of the twitter stream. ACM Transactions on the Web **9**(3), 12:1–12:33 (2015)
31. Zhou, L., Wang, W., Chen, K.: Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In: Proceedings of the 25th International Conference on World Wide Web (WWW'16) (2016)