Longitudinal Privacy Management in Social Media: The Need for Better Controls

This large-scale measurement study of Twitter focuses on understanding how users control the longitudinal exposure of their publicly shared social data — that is, their tweets — and the limitations of currently used control mechanisms. Our study finds that, while Twitter users widely employ longitudinal exposure control mechanisms, they face two fundamental problems. First, even when users delete their data or account, the current mechanisms leave significant traces of residual activity. Second, these mechanisms single out withdrawn tweets or accounts, attracting undesirable attention to them. To address both problems, an inactivity-based withdrawal scheme for improved longitudinal exposure control is explored.

Mainack Mondal and Johnnatan Messias
Max Planck Institute for Software Systems

Saptarshi Ghosh
IIT Kharagpur

Krishna P. Gummadi
Max Planck Institute for Software Systems

Aniket Kate
Purdue University

"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**[1]

The unprecedented sharing of personal, user-generated content on online social media sites such as Twitter and Facebook has spawned numerous privacy concerns for the sites' users. In fact, most users' privacy preferences for content sharing evolve over time because of changes in their sensitivity, the content's relevance, users' biographical status and relationships, and so on. To that end, in this article, we focus on a dimension of user privacy that's becoming increasingly challenging to manage: *longitudinal privacy*. This challenge refers to the difficulty in controlling the exposure of socially shared data over time. The problem becomes even more complex as more time passes and both the amount of shared content grows and new technologies emerge, such as archival, timeline-based searches that make it easier to access historical content shared under outdated privacy preferences.

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

1. Is there evidence for users changing their privacy preferences for content shared on social media sites 5–10 years in the past? If so, what's the

 1089-7801/17/$33.00 © 2017 IEEE

Prior Work in Longitudinal Privacy Control

Privacy notices and controls on social media sites have garnered considerable attention in recent times. Nevertheless, relatively little research explores longitudinal privacy management mechanisms. Two significant efforts are as follows. Oshrat Ayalon and Eran Toch find that a user's willingness to share content drops as the content becomes old; and that willingness drops further with a life-change event, such as graduating from college or moving to a new town.[1] Lujo Bauer and his colleagues discover that users want some old posts to become more private over time (and some to become more prominent), and their desired exposure for most content remained relatively constant over the years.[2] Both of these studies indicate that users are, in general, concerned about the privacy of their old content, providing a strong motivation for us to study at large scale how users in the real-world control their longitudinal privacy.

A natural way for users to protect their longitudinal privacy is to delete old content. In this direction, Hazim Almuhimedi and his colleagues reported the largest study so far on deleted tweets using real-world data; however, they collected only data deleted within a week after posting.[3] Specifically, they collected 67 million tweets from 292,000 users posted during a week, and found that 2.4 percent of those tweets were deleted within that week. Out of their set of deleted tweets, 89.1 percent were deleted on the same day of posting. Notice that they primarily focused on content posted in the near past (no more than one week old), which was selectively deleted by the user, while our article shows how the exposure controls are quite different for the content posted in the near and far past. As we demonstrate, that large study missed a considerable part of the deleted tweets, which were posted years before.

Oshrat Ayalon and Eron Toch[1] also propose longitudinal privacy management mechanisms, such as allowing users to set expiration dates on content or have an archive feature for old content. The advent and popularity of systems such as Snapchat (www.snapchat.com), which deletes all users' posts after a predefined expiry time, suggest users' enthusiasm for such age-based withdrawals. In this article, we demonstrate the limitations of age-based withdrawals, and propose a smarter mechanism that tries to decide dynamically which content to delete or archive based on its longitudinal exposure.

References

1. O. Ayalon and E. Toch, "Retrospective Privacy: Managing Longitudinal Privacy in Online Social Networks," *Proc. 9th Symp. Usable Privacy and Security*, 2013, article no. 4.
2. L. Bauer et al., "The Post Anachronism: The Temporal Dimension of Facebook Privacy," *Proc. 12th ACM Workshop on Privacy in the Electronic Society*, 2013, pp. 1–12.
3. H. Almuhimedi et al., "Tweets Are Forever: A Large-Scale Quantitative Analysis of Deleted Tweets," *Proc. 16th Conf. Computer Supported Cooperative Work*, 2013, pp. 897–908.

extent of the change in longitudinal exposure of user data?
2. How effective are the mechanisms that social media sites provide to enable users to control the exposure of their shared data over time? How could we improve the effectiveness of these longitudinal exposure control mechanisms?

To address these questions, we gathered extensive longitudinal data (over six years) from the Twitter social media site. We analyze these Twitter messages to seek answers to the aforementioned questions and suggest better longitudinal exposure control mechanisms for online social media sites. We conducted this study respecting the guidelines set by our institute's ethics board and with its explicit knowledge and permission.

Understanding Longitudinal Exposure

Our first task is to understand if and how users are presently withdrawing their socially shared content to control longitudinal exposure.

Collecting Data from Twitter

In Twitter, users can employ three distinct mechanisms to withdraw their content:

- selectively delete tweets,
- delete their entire user account, or
- make their account private.

If we query the Twitter API with a withdrawn tweet's previously archived tweet ID (the Twitter-generated unique identifier for a tweet), then Twitter returns an error code and message. From this error code and message, we can uniquely identify which mechanism was used to withdraw the tweet.[2] Some tweets are removed by Twitter when it suspends user accounts (such as spammer accounts); we ignore these tweets because they're withdrawn by Twitter and not the user.

We measured longitudinal exposure control of user data over the six-year period from July 2009 to October 2015. Specifically, we archived tweets from 2009 onward that were publicly posted at the time of collection. As Figure 1

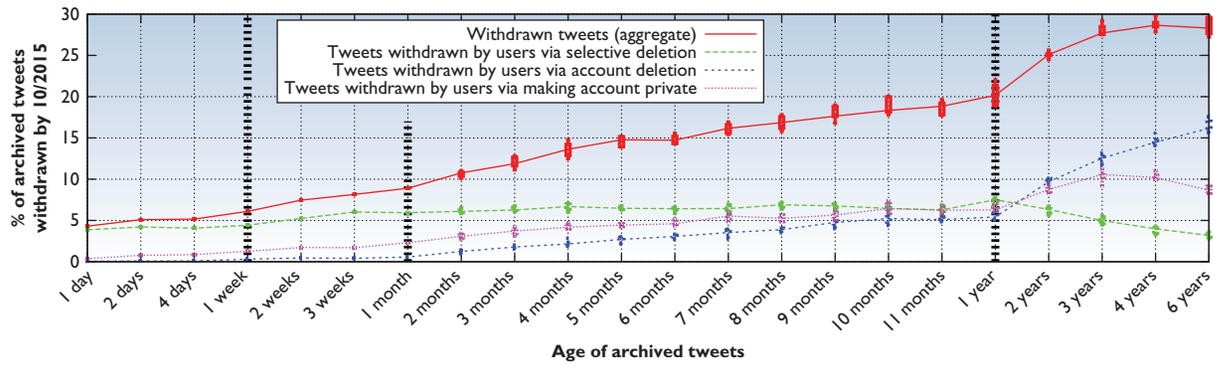

Figure 1. The percentage of tweets in our sample of archived tweets that were withdrawn as of October 2015. A tweet's age is the difference between the time the tweet was posted and the time that we queried the Twitter API with the tweet ID (October 2015). The amount of withdrawn tweets increased considerably over time — more than 28 percent of tweets posted six years before had been withdrawn. The dotted vertical lines demarcate the points on the x-axis where the scale changes (from days to months to years).

shows, on October 2015, we fixed 22 time periods over that six-year period, ranging from one day to six years. We randomly sampled 5,000 tweets from each timestamp, and used multiple samples for time periods for data older than two months; we then checked on October 2015 to see what fraction of tweets from these samples were withdrawn to measure the longitudinal exposure control.

Measuring Longitudinal Exposure Control in Twitter

How much of the archived data has been withdrawn? Figure 1 shows the variation in the percentage of tweets that were withdrawn for each time period. We show box and whiskers plots for time periods that are greater than or equal to two months, representing results from multiple days around those timestamps. We observe little variation among results from the repeated experiments over multiple consecutive days. Unless otherwise stated, we report here the median from the values obtained through the repeated experiments.

We discovered that a substantial amount of past data was withdrawn. As the solid red curve in Figure 1 shows, the percentage of withdrawn tweets increases from 4.3 percent of the tweets archived on the previous day (of our experiment) to 28.3 percent of the tweets archived in 2009. Hence, the natural next question is: How do the different exposure control mechanisms account for this data inaccessibility?

We then asked: What's the relative usage of different control mechanisms for longitudinal exposure? Figure 1 also shows the percentage of tweets withdrawn via the three longitudinal exposure controls:

- users selectively deleting tweets (green dashed curve),
- users deleting their account (blue curve), and
- users making their account private (pink curve).

Surprisingly, we found that tweets posted more recently compared to those posted in the distant past were withdrawn using very different exposure controls. Tweets posted more recently (such as one month prior to the last day) were mostly withdrawn by users selectively deleting their tweets. However, the percentage of tweets withdrawn via selective deletion quickly stabilizes over time. 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 (such as those six years back). Specifically, out of 8.9 percent of the withdrawn tweets from September 2015 (one month back), 5.9 percent were selectively deleted by users and only 3 percent were from users who deleted their account or made it private. In contrast, of the 28.3 percent withdrawn that were posted in 2009, as much as 16.2 percent were from users who deleted their account, while only 3.2 percent were selectively deleted. This global view motivated us to better understand privacy-related behaviors at a user level – that is, how are individual users controlling their longitudinal exposure?

Understanding User Behavior for Controlling Longitudinal Exposure

To assess individual users' behavior for controlling longitudinal exposure, we leverage a near-complete snapshot of Twitter data collected in September 2009.[3] 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 ones they posted (as obtained from the dataset). To simplify further analysis, we selected only English tweets and removed all suspended users and their tweets. We were left with 8,950,942 tweets (more than 89.5 percent of all tweets) posted by 97,998 users (more than 97.9 percent of all users).

Using the methodology described earlier, we found that 29.1 percent of all the tweets we checked were withdrawn during the six-year period; these tweets were posted by 34.6 percent of the selected users.

Based on longitudinal exposure control, we place our users into three distinct categories:

- *Non-withdrawers* — the 65.4 percent of users who did not withdraw any tweets.
- *Partial withdrawers* — the 8.3 percent of users who selectively withdrew some of their tweets, contributing 9.7 percent of the withdrawn tweets.
- *Complete withdrawers* — the 26.3 percent of users who withdrew all their tweets by either deleting their account or making their account private, contributing to 90.3 percent of all withdrawn tweets.

After understanding the privacy settings of different users and observing their significant use of longitudinal exposure controls, we investigate our next question: What are the limitations of the current exposure controls?

Limitations of Existing Longitudinal Exposure Controls

We observe that, across online social media sites, the existing longitudinal exposure control mechanisms have two inherent limitations: they retain *residual activities* associated with a withdrawn post (such as a deleted tweet) or a withdrawn (deleted or private) account, and they create signals to identify potentially sensitive content.

Limitation I: Retaining Residual Activities

On social media sites, users frequently engage in conversations with other users, spurring interactions linked to their posts or to their accounts — for example, by mentioning a user in a tweet or by tagging a user in a Facebook post, respectively. Even when users delete their posts or withdraw their entire account, these interactions remain in social media and become residual activities that continue to point to the withdrawn post or account. We observed that anyone today can collect a number of residual activities (that is, residual tweets on Twitter) around tweets and accounts withdrawn as far back as six years before the time of our study. Moreover, using these residual activities we can, in fact, retrieve information such as words or meanings within the withdrawn tweets, as well as demographics or interests of the withdrawn user accounts.

Next, we measure the volume of residual activities around withdrawn accounts and show how we use these activities to recover the interests of the withdrawn accounts. Detailed analysis and results are available elsewhere.[2]

Measuring residual activity around withdrawn accounts. Two widely employed longitudinal exposure control mechanisms in Twitter are to make an account private or to delete it. We considered these withdrawn accounts from our random sample of 97,998 users from 2009, and used the Twitter search to collect posts that mention any of those accounts. We limited our search to the period when the withdrawn accounts were active in our dataset — that is, from the account's creation date to the date of the last tweet appearing in our data. We collected a total of 1,403,716 residual tweets that mentioned 23,526 withdrawn accounts. In other words, as much as 91.4 percent of the withdrawn accounts had some residual tweets around them. Moreover, 55.9 percent of all withdrawn accounts had 10 or more residual tweets.

Recovering information from residual activities. We found that, by leveraging residual activities around withdrawn tweets, we not only found keywords from those withdrawn tweets, but also recovered their meanings. The situation for withdrawn accounts is even worse — we can recover social connections, demographics, and even the interests of the withdrawn accounts by leveraging the residual activities.

Recovering interests of withdrawn accounts. To concretely demonstrate the problem with residual activities, we now describe the recovery of

Table 1. Hashtags revealed by residual tweets for 10 withdrawn accounts.

User number	Topics	Hashtags used by withdrawn accounts (revealed by residual tweets)
1	Politics, sports, technology	#iranelection, #prisoners, #strike, #frenchopen, #tech
2	Politics	#conservativebabesarehot, #teaparty, #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

interests of withdrawn accounts.[2] We leveraged a special type of keyword — hashtags — to identify potential topics of interest for withdrawn accounts. Hashtags are words in tweets that start with a "#" symbol and are included to provide the tweet a specific context and information about the interests of users. We observed that 3,855 accounts in our set of withdrawn accounts posted at least one tweet with a hashtag. Of those, the residual tweets revealed at least one hashtag for 58.7 percent of accounts (2,263 in total) and a total of 3,625 unique hashtags. Interestingly, for 25 percent of the withdrawn accounts that used hashtags, *all* the hashtags revealed by the residual tweets were also used in the tweets posted by the withdrawn accounts.

Table 1 shows 10 individual withdrawn accounts and the hashtags revealed from residual tweets for those accounts. We verified that the withdrawn account owners themselves actually used each of these hashtags. The table also shows some manually annotated topical categories for these hashtags. The hashtags give us an idea of what topics might interest the owners of the withdrawn accounts.

We also manually classified the hashtags into topics. Interestingly, some of these hashtags, such as "#iranelection" and "#nsfw," might be considered sensitive, while others, such as "#daviscup," "#tech," and "#nascar" convey specific interests of the withdrawn accounts.

Twitter app to raise awareness about residual activities. To increase user awareness about their residual activities, we designed a Twitter app that lets Twitter users check the information about their account and individual tweets that can be inferred by simply analyzing their residual Twitter activities. The app is available at http://twitter-app.mpi-sws.org/footprint and we encourage you to use it.

Limitation II: Creating Signals to Identify Potentially Sensitive Content

In addition to social media operators, the content posted on social media is also collected and stored by a plethora of third parties for reasons ranging from simple archiving to nefarious attempts to mine sensitive user information. Although following a few high-profile users to look for sensitive content is easy for these miners, it's difficult to scavenge the astronomical volume of daily user-generated data to find sensitive information from unspecified targets to cyberstalk them later.

However, current longitudinal access control mechanisms provide an easy way to single out sensitive information about a user. The intuition is simple: If a user wants to withdraw some old content, it might well contain personal or embarrassing information and merit further analysis. This limitation isn't specific to Twitter, but rather holds for any site that lets users delete their records. A curious analyst can collect publicly accessible data at any point of time, archive it, and later check which data records are deleted to identify personal or embarrassing information. Moreover, this scenario of attracting more attention to historic content when users try to control its exposure is increasingly common in the real world. Following are two examples.

Deleted tweets from politicians. Twitter allows third-party analysts to collect tweets in real time.

For each tweet sent to an analyst, Twitter sends a *deletion notification* if the original user deletes that tweet. Although Twitter's intention is to alert analysts so that they can delete that tweet from their database, a malicious analyst can quite easily use these notifications to single out deleted tweets. The Politwoops service (http://politwoops.sunlightfoundation.com) actually leverages these features to identify and publish deleted tweets by politicians.

YouTube deleted videos. Even in the absence of deletion notifications, analysts can simply check which of the social media posts they archived have been withdrawn by making http requests to the site. The YouTomb service (https://en.wikipedia.org/wiki/YouTomb) used this strategy to build a searchable database of recent video removals on YouTube. The website is unavailable as of November 2014.

Toward Better Longitudinal Exposure Control

As the previous examples demonstrate, it's inherently difficult to protect sensitive information withdrawals from prying eyes. Generally, both limitations on longitudinal exposure control are fundamental to how these mechanisms are expected to work today: residual activities are owned by their creators and can't be withdrawn proactively, while content deletion notices are essential for withdrawal of replicated user-generated content.

Longitudinal exposure control is clearly a complex problem; the fundamental nature of its limitations demonstrates that indisputably. Although it's highly unlikely that a silver bullet will solve all the problems with longitudinal exposure control, we propose alternative mechanisms that can be highly effective for social media sites such as Twitter.

Content Anonymization

Users typically withdraw past content for two reasons: the content is sensitive (for example, it contains swear words) or the content isn't sensitive in itself, but users don't want it attached to their identity. We propose a simple mechanism for the second case: Social network operators give users the option of anonymizing the content. If users choose this option, all possible traces of their identity are removed from the content and any directly associated residual activities. In Twitter, as a first step, this can be realized by simply removing all user mentions from a withdrawn tweet or account and the residual activities pointing at it. This involves removing @mentions, withdrawn tweet-IDs, or withdrawn account-IDs. As of 2016, another social media operator, Reddit, began employing a basic form of this type of anonymization by unlinking all posts from withdrawn accounts (www.reddit.com/wiki/privacypolicy).

Addressing the residual activities limitation. Content anonymization solves the problem associated with residual activities because, after anonymization, it's impossible to link a residual activity to a particular user or post or to establish a link between multiple residual activities. As a result, analysts scavenging for residual activities become severely restricted in their ability to infer identifying information about withdrawn content.

Addressing the withdrawn content signal limitation. Content anonymization doesn't delete posted content, it simply unlinks the posting user from the content. A resourceful analyst can still periodically fetch all the content and do a *diff* (comparing text to find the difference) to determine if the content is anonymized. Thus, in principle, it's still possible for analysts to determine if content is withdrawn. As a result, content anonymization doesn't address the limitation related to creating a withdrawn content signal.

Other issues. Content anonymization has a practical problem: although simply removing all pointers and identifiers is a good first step, personally identifiable information (PII) can persist in the content in the form of the name of the user or close associates, or even in the form of writing style. Removing PII is an active research area that, while promising, might not be robust enough to deploy yet. These problems bring us to our next, more drastic proposal.

Inactivity-Based Withdrawal

Ephemeral social media sites such as Snapchat (www.snapchat.com) and Cyber Dust (www.cyberdust.com) offer users *age-based withdrawal* to control longitudinal exposure. On such sites, every message is associated with an expiry time, after which the post is automatically withdrawn and becomes inaccessible to users.

Age-based withdrawal has two limitations. First, the default expiry time is generally too small (such as a few seconds or minutes), which could prevent any meaningful discussion around posts. Second, users are generally poor at anticipating when a post should be deleted, which reduces the mechanism's practical use even if users are given the option of setting the expiry time.[4]

To that end, we propose a novel mechanism, *inactivity-based withdrawal*. Our proposal is based on a rather simple intuition: when a post becomes inactive — that is, when it doesn't generate any further interaction or receive any further exposure — it can be safely withdrawn (deleted/archived/hidden) from the public domain. Here, "interaction" is a general term; it might involve various tasks, depending on the social media site. For example, it might mean sharing the post (such as retweeting in Twitter), replying to the post, or even viewing the post on the original posting account or on the accounts of other users.

Compared to age-based withdrawal, this mechanism has the following advantages. First, users need not be burdened with deciding their posts' expiry times. Instead, the social site operator can present suggestions to users when a post becomes inactive and thereby facilitate its withdrawal. Second, it allows meaningful discussions around interesting posts, as posts are withdrawn only after discussion around them has died down. Inactivity-based withdrawal also overcomes both limitations with the current longitudinal exposure control mechanism as the withdrawals (both the original post and residual activities) are automatic, so there's generally no motivation for analysts to seek out withdrawn content or the withdrawer's account.

A technical question, however, must be addressed: How do we select a time period T of inactivity after which a post will be withdrawn? As we evaluated earlier,[2] a social site's operator has enough information to do it reasonably well; in particular, operators can leverage past interaction history to select an appropriate T value (such as 180 days or six months) that stops only a fraction of interactions. In fact, the system operator could show a range of threshold values and point out the associated percent of stopped activities based on a user's past history, then let each user make an informed decision.

However, some users might still want to withdraw sensitive content while it's still generating interaction or before it reaches its expiry age determined by T. Analysts with auxiliary knowledge about the generated interaction and the T value might be able to detect such incidents; however, the system operator can easily keep most of the generated interaction invisible from analysts and can also pick the T value judiciously to significantly reduce this success probability.

Although we don't claim that our proposed mechanism solves all the problems with longitudinal exposure control, it's a step toward more usable longitudinal exposure control mechanisms.

Although our study uses Twitter data, the phenomenon of content withdrawal is widespread on multiple existing social media sites today.[5] Moreover, the problems with existing privacy mechanisms — such as residual activity — aren't specific to Twitter. For example, Facebook posts by other users mentioning names of deleted accounts are also residual activities and can lead to problems similar to those we discuss here. To conclude, our study also calls for further research in this field, since much remains to be done in this space of understanding and improving longitudinal exposure controls of socially shared data.

References

1. H.W. Jenkins Jr., "Google and the Search for the Future," *The Wall Street J.*, 14 Aug. 2010; www.wsj.com/articles/SB10001424052748704901104575423294099527212.
2. M. Mondal et al., "Forgetting in Social Media: Understanding and Controlling Longitudinal Exposure of Socially Shared Data," *Proc. 12th Symp. Usable Privacy and Security*, 2016; www.usenix.org/conference/soups2016/technical-sessions/presentation/mondal.
3. M. Cha et al., "Measuring User Influence in Twitter: The Million Follower Fallacy," *Proc. 4th AAAI Conf. Weblogs and Social Media*, 2010, pp. 10–17.
4. L. Bauer et al., "The Post Anachronism: The Temporal Dimension of Facebook Privacy," *Proc. 12th ACM Workshop on Privacy in the Electronic Society*, 2013, pp. 1–12.
5. M. Madden et al., *Teens, Social Media, and Privacy*, Pew Research Center, 2013; www.pewinternet.org/2013/05/21/teens-social-media-and-privacy.

Mainack Mondal is a doctoral student at the Max Planck Institute for Software Systems in Germany. His research interests are in networked systems, with an emphasis on user privacy. Mondal has an MTech in computer science

and engineering from the Indian Institute of Technology, Kharagpur. Contact him at mainack@mpi-sws.org.

Johnnatan Messias is a master's candidate at Universidade Federal de Minas Gerais, Brazil; he was previously a research intern at Max Planck Institute for Software Systems. His research interests include data science, social computing, mobile computing, bots, and autonomous flight with drones. Messias has a bachelor's degree in computer science from the Universidade Federal de Ouro Preto, Brazil. Contact him at johnme@mpi-sws.org.

Saptarshi Ghosh is an assistant professor in the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, and was a Humboldt postdoctoral fellow at the Max Planck Institute for Software Systems. His research interests include social media, network analysis, information retrieval, and data mining. Ghosh has a PhD in computer science and engineering from the Indian Institute of Technology Kharagpur. Contact him at saptarshi@cse.iitkgp.ernet.in.

Krishna P. Gummadi is a tenured faculty member and head of the Networked Systems Research Group at the Max Planck Institute for Software Systems in Germany. His research interests include the measurement, analysis, design, and evaluation of complex Internet-scale systems, as well as understanding and building social computing systems. Gummadi has a PhD in computer science and engineering from the University of Washington. Contact him at gummadi@mpi-sws.org.

Aniket Kate is an assistant professor in the Computer Science Department at Purdue University; he previously completed a postdoctoral fellowship at Max Planck Institute for Software Systems. His research interests include designing, implementing, and analyzing transparency- and privacy-enhancing technologies, integrating cryptography, distributed systems, and hardware-assisted security. Kate has a PhD in computer science and engineering from the University of Waterloo, Canada. Contact him at aniket@purdue.edu.