Editors: Heather Richter Lipford, richter@uncc.edu | Jessica Staddon, jessica.staddon@gmail.com

User Anonymity on Twitter

Sai Teja Peddinti | Google
Keith W. Ross and Justin Cappos | New York University

The Internet's proliferation has resulted in a growing number of online social networks and discussion forums. To participate, users must typically create an account and adopt an online identity. Services often differ in their acceptable user identity requirements. For example, Facebook enforces a *real-name* policy, requiring users to supply their real names when creating accounts. Stated reasons for this include that such a policy increases user accountability and improves content quality (by helping decrease spam, bullying, and hacking). However, privacy advocates claim that real-name policies erode online freedom by letting services tie user interests (as reflected by their online actions) to their names, thereby generating a treasure trove of information.[1]

Twitter, on the other hand, doesn't require users to provide their real names, although it does ask them to create unique pseudonyms. Using pseudonyms with no relation to their real names effectively makes users anonymous (that is, anonymous to other users of the service, though not necessarily to the service provider). The absence of a real-name policy has made Twitter a popular information exchange portal for users to share and access information without being identifiable.[2,3]

Online and offline anonymity have both been extensively studied;[4–6] here, we focus specifically on how this anonymity influences user behavior in online social networks. We conducted a large-scale, data-driven analysis of Twitter to identify the prevalence of user anonymity and its correlation with content sensitivity. (To learn more about the three Twitter datasets we used, see the sidebar.) We also explored the feasibility of an automated system that leverages user anonymity patterns to help identify sensitive content. Through our work, we hope to develop a deeper understanding of anonymity's importance and role in society, guide the development of new privacy and anonymity features in existing and future online social networks, and discover potentially sensitive or controversial topics in social networks. For ease of reading, we'll use the commonly used term *anonymous* here rather than the more obscure *pseudonymous*.

Twitter Account Basics

Every Twitter account contains four main pieces of information:

- A *profile* in which the user provides personal details, including a unique alphanumeric ID identifying the account, known as a screen name; a name field, which usually contains the user's first and last name; a profile picture; and a URL, which might link to another social network profile. Note that the details provided in the profile need not always be true; for example, the name field could consist of a fake first name, fake last name, or both.
- A list of the *tweets*, or messages, posted by the user.
- A *friends* list. When a user follows another user, or "friend," he or she receives tweet updates from that friend. This relationship is unidirectional: if Alice is a friend of Bob, Bob need not be a friend of Alice.
- A *followers* list. The other users who receive all of the user's tweet updates are termed "followers."

Our Work

To measure the prevalence of anonymity in Twitter, we randomly picked 100,000 accounts from a 2010 public Twitter dataset containing 41.7 million accounts.[7] After eliminating all the deactivated accounts, non-English accounts (those not reporting English as the language of preference), spam accounts, and inactive/ephemeral accounts, we passed a dataset of 50,173 Twitter accounts to Amazon Mechanical Turk (AMT) for labeling.

Copublished by the IEEE Computer and Reliability Societies

Labeling Twitter Accounts

We supplied each Twitter account's name and screen name to AMT workers, asking them to determine whether these two fields collectively contained just a first name, just a last name, both a first and a last name, or neither a first nor a last name. Workers could also indicate that they were unsure. On the basis of this AMT labeling, we assigned each account to one of the following categories:

- *anonymous*—a Twitter account with neither a first nor a last name and no URL in the profile (because a URL could point to a webpage that partially or fully identified the user).
- *partially anonymous*—a Twitter account with either a first or a last name but not both.
- *identifiable*—a Twitter account with both a first and a last name; or
- *unclassifiable*—any Twitter account not falling into one of the above categories, such as accounts with a URL but no first or last name, or organizational or company Twitter accounts.

Note that noise in user classification is difficult to completely remove. For instance, a small fraction of the accounts labeled anonymous might not have been fully so in that the users provided an identifiable profile photo or disclosed their identities in their tweets. Furthermore, a fraction of identifiable accounts might have been effectively anonymous because the users provided fake first and last names.

Quantifying User Anonymity

We found that 6 percent of the analyzed accounts were anonymous, as they didn't disclose a first or last name. Another 20 percent were partially anonymous, disclosing only a first or a last name. This signifies that online anonymity is important to at least one-quarter of the Twitter population, and that Twitter's lack of a real-name policy could be a strong selling point for the social network. Of the remaining accounts, 6 percent were unclassifiable and 68 percent were identifiable. Of course, some of the identifiable users might have been using fake first and last names and thus were actually anonymous. This implies that the 26 percent of users categorized as not fully disclosing their identity on Twitter was likely a low estimate.

User Anonymity and Content Sensitivity

To evaluate whether content sensitivity correlates with users choosing to be anonymous, we selected several broad topic categories widely considered to be sensitive and/or controversial: pornography, escort services, sexual orientation, religious and racial hatred, online drugs, and guns. For comparison, we also chose several nonsensitive categories: news sites, family recreation, movies/theater, kids/babies, and companies/organizations producing household items. For each category, we identified a few distinctive search terms and manually picked Twitter accounts that showed up when we searched those terms on the Twitter page.

We picked 50 Twitter accounts related to the sensitive categories, and 20 related to the nonsensitive. Figure 1 shows the average percentage of followers who were anonymous versus identifiable for each sensitive and nonsensitive category. The categories are arranged from highest to lowest percentage of anonymous followers.

The sensitive categories had the largest percentage of anonymous users: at least 21.6 percent of users following pornography, marijuana, Islamophobia, and gay/lesbian accounts were anonymous, with

Our Twitter Datasets

To achieve our research goals,[1,2] we studied three different Twitter datasets.

Our first goal was to measure how many users adopted anonymous pseudonyms. Determining this required a random sample of Twitter accounts. So we turned to the large Twitter crawl publicly available (as of early 2014), which was published in 2010.

Our second goal was to measure the correlation between content sensitivity and user anonymity. For this, we manually selected 70 sensitive and nonsensitive accounts in 2014 and studied their followers.

Finally, to determine whether we could build automated classifiers that would detect sensitive Twitter accounts, we later crawled Twitter and obtained 100,000 accounts and captured their 400 million followers.

We studied three datasets at different points over a span of five years. Not only did anonymous Twitter accounts exist in all three datasets, but the relationship between anonymous accounts and sensitive Twitter accounts didn't change across the different datasets. This gives us confidence in our results' validity. Moreover, future repetitions of the study could allow our methodology to identify new and emerging sensitive themes in addition to what's already been discovered.

References

1. "'On the Internet, Nobody Knows You're a Dog': A Twitter Case Study of Anonymity in Social Networks," *Proc. ACM Conf. Online Social Networks* (COSN 14), 2014, pp. 83–94.
2. "Finding Sensitive Accounts on Twitter: An Automated Approach Based on Follower Anonymity," *Proc. Int'l AAAI Conf. Web and Social Media* (ICWSM 16), 2016, pp. 665–658.

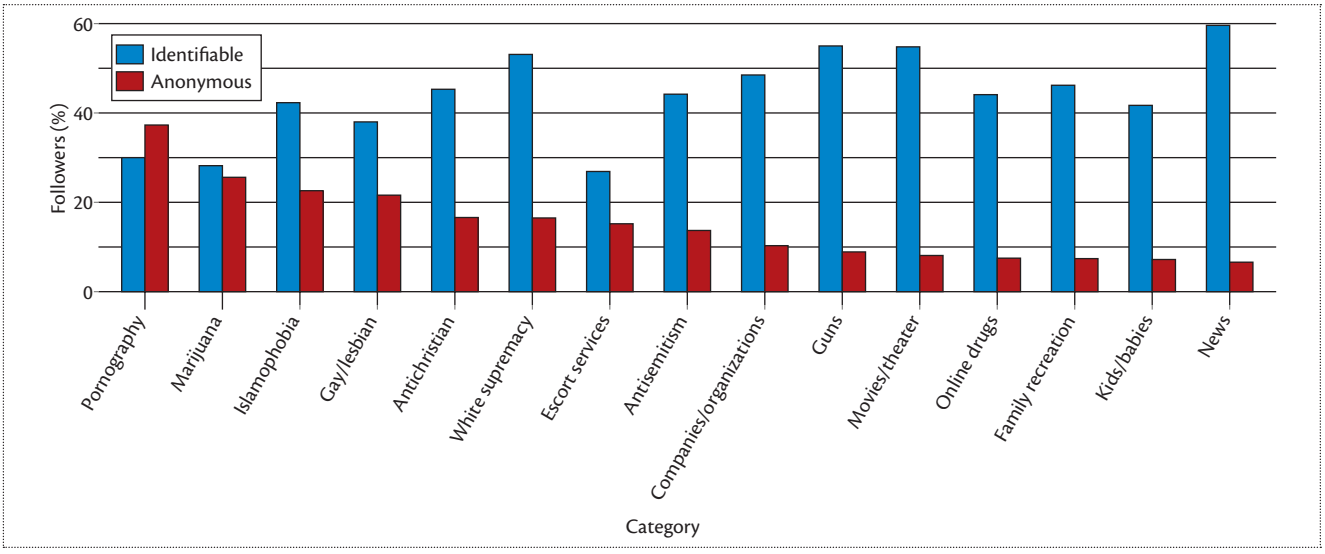

Figure 1. Sensitive and nonsensitive Twitter account categories, arranged from highest to lowest percentage of anonymous followers.

pornography far exceeding the rest with 37.3 percent anonymous followers. However, some sensitive categories—such as white supremacy and guns—had a surprisingly large percentage of identifiable followers. It appears that some types of sensitive content generate secrecy, while others encourage openness. This observation reaffirms that content sensitivity is quite nuanced and complex.

Even nonsensitive categories have 6.6 to 8.9 percent anonymous followers. This observation confirms that users don't create anonymous profiles for the sole purpose of following sensitive accounts. To avoid maintaining multiple profiles, an anonymous user might follow both sensitive and nonsensitive accounts using the same profile, leaking out his or her interests on Twitter.

Automatically Detecting Sensitive Accounts

One way to identify sensitive accounts is to specify categories of sensitive topics, identify words that commonly occur when discussing these topics, and then search for tweets and accounts that employ these words. However, this approach is highly subjective because it relies on humans to define the sensitive topics and words.

Another approach is to apply automated topic identification techniques, such as latent Dirichlet allocation (LDA), to tweets. This allows identification of accounts related to these sensitive themes. However, such techniques are highly resource intensive and can't scale to Twitter's size.[8]

So, we investigated whether our observed user anonymity patterns and their correlation to content sensitivity could be leveraged to develop an efficient, automated means of identifying accounts that tweet sensitive content. Such an approach would generalize better to unforeseen topics, wouldn't be limited by language features, and would be easily scalable.

We first considered the subproblem of automatically determining whether a Twitter account was anonymous or identifiable. We relied on the previously labeled Twitter accounts for training. Because anonymous and identifiable accounts differ in the presence of first and last names, we captured the US Census's and Social Security Administration's public first and last name lists.[5] However, simply checking for occurrences in the name lists resulted in very poor anonymous and identifiable detection rates. So we extracted additional information available from Twitter profiles such as popularity ranks of the occurring first and last names in the public name lists; name strings following structural constraints (such as "FirstName MiddleInitial LastName"); and number of friends, followers, tweets, and so on.

Using these extracted features, we trained a Random Forests–based anonymity machine learning classifier that can accurately detect anonymous and identifiable accounts with more than 90 percent precision. Then, on the basis of the fraction of anonymous and identifiable followers detected by our anonymity classifier across the known 70 sensitive and nonsensitive accounts studied earlier, we developed a Support Vector Machine–based sensitivity classifier that can separate sensitive and nonsensitive Twitter accounts.

To test our sensitivity classifier, we crawled Twitter and captured a random sample of 100,000 accounts with approximately 404 million active followers. We applied our classifier to these accounts after

labeling their followers as either anonymous or identifiable.

Manual inspection showed that the top accounts determined to be sensitive by our classifier were indeed discussing topics that many would consider sensitive: pornography, drugs, and adult content. However, in addition to these usual suspects, our approach uncovered many accounts related to socially desirable themes, emphasizing that anonymity serves many ends.

For example, we identified many accounts supporting and fighting for lesbian, gay, bisexual, transgender, and queer rights. Disclosing one's sexual orientation is a sensitive issue for many, and hence users might prefer not to identify themselves. We found accounts where users openly discuss marital and other relationship issues, share personal feelings or experiences, and address health issues. Anonymity might offer an opportunity for people to solicit support or find solace.

We also discovered accounts dealing with severe cases of anorexia, social anxiety, depression, and suicidal tendencies. In fact, on some of these accounts, the users uploaded pictures after having physically abused their bodies. While these accounts have varied aims, health institutions are using them as inroads for reaching out to people who might need help.[9]

The existence of accounts related to these sensitive themes—and the fact that they have many anonymous followers—supports the thesis that privacy and anonymity are important in our society.

Although our novel methodology for identifying sensitive accounts on Twitter provides a scalable and objective way to understand content sensitivity, more in-depth research is needed to improve user privacy preferences and expectations in the social media context. For instance, it's worth exploring and quantifying how many sensitive content categories are consistent across different social applications and how many depend on the application's nature (such as photo sharing versus messaging). We hope our findings will contribute to future improvements in privacy policies and new privacy controls. ■

Acknowledgments

This article is based on the authors' two prior publications, "'On the Internet, Nobody Knows You're a Dog': A Twitter Case Study of Anonymity in Social Networks" (*Proc. ACM Conf. Online Social Networks* [COSN 14], 2014, pp. 83–94), and "Finding Sensitive Accounts on Twitter: An Automated Approach Based on Follower Anonymity" (*Proc. Int'l AAAI Conf. Web and Social Media* [ICWSM 16], 2016, pp. 665–658).

References

1. N. Lomas, "Facebook Users Must Be Allowed to Use Pseudonyms, Says German Privacy Regulator; Real-Name Policy 'Erodes Online Freedoms,'" Techcrunch, 18 Dec. 2012; techcrunch.com/2012/12/18/facebook-users-must-be-allowed-to-use-pseudonyms-says-german-privacy-regulator-real-name-policy-erodes-online-freedoms.
2. A. Kavanaugh et al., "Microblogging in Crisis Situations: Mass Protests in Iran, Tunisia, Egypt," *Proc. Workshop Transnational Human-Computer Interaction* (CHI 11), 2011; eventsarchive.org/sites/default/files/Twitter%20Use%20in%20Iran%20Tunisia%20Egypt.Kavanaugh.Final__0.pdf.
3. E. Mustafaraj et al., "Hiding in Plain Sight: A Tale of Trust and Mistrust inside a Community of Citizen Reporters," *Proc. 6th Int'l AAAI Conf. Weblogs and Social Media* (ICWSM 12), 2012, pp. 250–257.
4. M.S. Bernstein et al., "4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community," *Proc. 5th Int'l AAAI Conf. Weblogs and Social Media* (ICWSM 11), 2011, pp. 50–57.
5. D. Correa et al., "The Many Shades of Anonymity: Characterizing Anonymous Social Media Content," *Proc. 9th Int'l AAAI Conf. Web and Social Media* (ICWSM 15), 2015; socialnetworks.mpi-sws.org/papers/anonymity_shades.pdf.
6. S.T. Peddinti et al., "Cloak and Swagger: Understanding Data Sensitivity through the Lens of User Anonymity," *Proc. 35th IEEE Symp. Security and Privacy*, 2014, pp. 493–508.
7. H. Kwak et al., "What Is Twitter, a Social Network or a News Media?," *Proc. 19th Int'l Conf. World Wide Web* (WWW 10), 2010, pp. 591–600.
8. B. Bi et al., "Scalable Topic-Specific Influence Analysis on Microblogs," *Proc. 7th ACM Int'l Conf. Web Search and Data Mining* (WSDM 14), 2014, pp. 513–522.
9. J. Jashinsky et al., "Tracking Suicide Risk Factors through Twitter in the US," *Crisis*, vol. 35, no. 1, 2014, pp. 51–59.

Sai Teja Peddinti is a research scientist in the Security and Privacy group at Google. This research was done while he was a PhD candidate at New York University (NYU). Contact him at psaiteja@google.com.

Keith W. Ross is the dean of Engineering and Computer Science at NYU Shanghai and the Leonard J. Shustek Chair Professor of Computer Science and Engineering at NYU. Contact him at keithwross@nyu.edu.

Justin Cappos is an assistant professor in the Tandon School of Engineering at NYU. Contact him at jcappos@nyu.edu.