"On the Internet, Nobody Knows You're a Dog": A Twitter Case Study of Anonymity in Social Networks

Sai Teja Peddinti[*]
psaiteja@nyu.edu

Keith W. Ross[*†]
keithwross@nyu.edu

Justin Cappos[*]
jcappos@nyu.edu

[*]Dept. of Computer Science and Engineering, NYU
Brooklyn, New York, USA

[†]NYU Shanghai
Shanghai, China

ABSTRACT

Twitter does not impose a Real-Name policy for usernames, giving users the freedom to choose how they want to be identified. This results in some users being *Identifiable* (disclosing their full name) and some being *Anonymous* (disclosing neither their first nor last name).

In this work we perform a large-scale analysis of Twitter to study the prevalence and behavior of Anonymous and Identifiable users. We employ Amazon Mechanical Turk (AMT) to classify Twitter users as Highly Identifiable, Identifiable, Partially Anonymous, and Anonymous. We find that a significant fraction of accounts are Anonymous or Partially Anonymous, demonstrating the importance of Anonymity in Twitter. We then select several broad topic categories that are widely considered sensitive–including pornography, escort services, sexual orientation, religious and racial hatred, online drugs, and guns–and find that there is a correlation between content sensitivity and a user's choice to be anonymous. Finally, we find that Anonymous users are generally less inhibited to be active participants, as they tweet more, lurk less, follow more accounts, and are more willing to expose their activity to the general public. To our knowledge, this is the first paper to conduct a large-scale data-driven analysis of user anonymity in online social networks.

Categories and Subject Descriptors

J.4 [**Social And Behavioral Sciences**]: Sociology; K.4.1 [**Public Policy Issues**]: Privacy; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement, Human Factors

Keywords

Online Social Networks; Twitter; Anonymity; Quantify; Behavioral Analysis

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
COSN'14, October 1–2, 2014, Dublin, Ireland.
Copyright 2014 ACM 978-1-4503-3198-2/14/10 ...$15.00.
http://dx.doi.org/10.1145/2660460.2660467.

1. INTRODUCTION

Many online social networks, including Facebook and Google+, enforce a Real-Name policy, requiring users to use their real names when creating accounts [3, 2]. The cited reasons for the Real-Name policy include that it improves the quality of the content and the service (helping decrease spam, bullying, and hacking), increases accountability, and helps people to find each other. The Real-Name policy, however, also enables the social networks to tie user interests–as reflected from their use of the online services–with their true names, generating a treasure trove of consumer data. This has resulted in many debates [13] and petitions [6], with privacy advocates claiming that Real-Name policy erodes online freedom [31]. Privacy-conscious users have started finding ways to bypass the policy, hiding their real identity while continuing to use these social networks [22].

Twitter, on the other hand, does not impose strict rules for users to provide their real names, although it does require them to register with and employ unique pseudonyms. Taking advantage of this lack of Real-Name policy, many Twitter users choose to employ pseudonyms that have no relation to their real names. Some users choose such a pseudonym only because they enjoy being associated with a particular fun or interesting pseudonym. But many users likely choose pseudonyms with no relation to their real names because they want to be anonymous on Twitter. For example some users may desire the ability to tweet messages without revealing their actual identities. Other users may desire to follow sensitive and controversial accounts without exposing their real identities. The lack of Real-Name policy enforcement has turned Twitter into a popular information exchange portal where users share and access information without being identifiable–as is evident by Twitter's role in Egyptian revolution [25] and for reporting news in Mexico [34]. However, there is a meaningful debate about the pros and cons of online anonymity, as it allows people to more easily spread false rumours [14], defame individuals [12], attack organizations [33], and even spread spam [41, 17].

In this work we use Twitter to study the *prevalence* and *behavior* of Identifiable users (those disclosing their full name) and Anonymous users (those disclosing neither their first nor last name). Although both on-line and off-line anonymity has been considered by researchers in psychology and sociology, as discussed in Section 7, these studies have generally been carried out with small data sets and surveys. There have also been a few data-driven studies of anonymity in blogs and postings to Web sites [16, 5, 36]. To our

knowledge, this paper is the first to conduct a large-scale data-driven analysis of user anonymity in online social networks. The potential benefits of such a study include: (i) a deeper understanding of the importance and role of anonymity in our society; (ii) guidance for the incorporation of privacy and anonymity features in existing and future online social networks; (iii) and as we shall discuss in the body of the paper, the discovery of illegal (such as child-porn and terrorism) or controversial (such as ethnic or religious hate) activities.

Contributions

- We first analyze a large random sample of 100,000 Twitter users. After removing ephemeral users (active on Twitter for less than six months) and spam users, we employ Amazon Mechanical Turk (AMT) to classify Twitter users as Highly Identifiable, Identifiable, Partially Anonymous, and Anonymous based on whether their first and last names are given in their profiles and whether they link to other social networks with a Real-Name policy. We find that 5.9% of the accounts are Anonymous and 20% of the accounts are Partially Anonymous, demonstrating the importance of Anonymity for a large fraction of Twitter Users. Leveraging this same data set, we find Identifiable and Anonymous users exhibit distinctly different behavior in choosing which accounts to follow.

- We evaluate whether content sensitivity has any correlation with users choosing to be anonymous. For this analysis we select several broad topic categories that are widely considered sensitive and/or controversial– pornography, escort services, sexual orientation, religious and racial hatred, online drugs, and guns. We also consider several generic non-sensitive categories. For each of these broad categories we identify Twitter accounts that tweet about these categories. We observe that the different categories contain greatly different percentages of Anonymous and Identifiable followers. Strikingly, all but one of the sensitive aggregate categories have the largest percentage of Anonymous users. We also examine each of the non-sensitive and sensitive accounts individually and observe that there is a general pattern of having larger percentages of Anonymous followers for the sensitive accounts and larger percentages of Identifiable followers for the non-sensitive accounts. As we discuss in the body of the paper, this observation can potentially lead to a new mechanism for identifying sensitive and controversial accounts, as well as helping to determine what types of categories people consider to be sensitive.

- We combine the two datasets and analyze some of the behavioral issues associated with Anonymous and Identifiable users. We find that Anonymous users are generally less inhibited to be active participants, as they tweet more, lurk less, follow more accounts, and are more willing to expose their activity to the general public. However, the Highly Identifiable users, who publicly link to OSNs with a Real-Name policy, typically have many more friends and followers than Identifiable users, demonstrating a high degree of online social activity and visibility.

The following sections of the paper are organized as follows. Section 2 provides a brief background on Twitter and its terminology. Section 3 gives details about the user categories we are interested in and the classification procedure. We describe our collected dataset statistics in Section 4. Our findings on the use of non-identifying pseudonyms, correlation with following sensitive accounts, and group behavioral differences are reported in Section 5. Section 6 discusses future work. Section 7 describes the related work and Section 8 concludes the paper.

2. BACKGROUND

Every Twitter account is comprised of four main pieces of information.

- First is the account *Profile* which includes the details provided by the user about him/her. These include the *screen name*, which is a user-chosen unique alphanumeric ID (also referred to as the username); the *name*, which may be the user's actual first and last name; and (optionally) a small textual description, a profile picture, the user's city/location and a URL (either linking to another social network profile or to something the user supports). It is to be noted that the details provided in the profile need not always be true (e.g., the name field can contain a fake first and/or last name).

- Second is the list of *Tweets* (i.e., messages) posted by the user. A tweet is a message restricted to 140 characters and can contain text, URLs (URL shortening is generally applied to limit the URL size to 20 characters) and *HashTags* (which is a metadata tag used to group messages).

- Third is the *Friends* list of the user. When a Twitter user follows another user (a "friend"), he/she receives the tweets from that friend. This relationship is unidirectional, so if A is a friend of B, B need not be a friend of A.

- Fourth is the *Followers* list of the user. All the users who follow a particular Twitter user are termed his/her followers. They receive all the tweet updates posted by the particular user.

By default, all of this information is publicly available from the Twitter web site. Twitter provides a *protected* privacy feature, to enable users to hide their tweets, friend lists, and follower lists.

Twitter provides a free API to obtain nearly unrestricted access to the social network data, which is only limited by the number of requests that can be sent during a time interval. In this work, we limit our analysis to the profile information, friends and followers listing and do not analyze the tweets posted by the user.

Ephemeral and Spam Accounts

In order to not bias the results, we remove from our data sets all user accounts that show signs of being ephemeral or spam. We say an account is *non-ephemeral* if the sum of friends and followers is at least five *and* it has had some activity–either (i) posting a tweet or (ii) adding a friend–at least six months after its creation. As the API doesn't give the dates that friends are added, we take a conservative approach for meeting condition (ii). For a given account Bob,

we examine the account creation dates of all the friends of Bob. If Bob has at least one friend with an account creation date that is six months after Bob's account creation date, then Bob clearly added a friend at least six months after creating his account.

Various entities frequently attempt to create spam accounts in Twitter for spreading spam or malware [17, 41]. Twitter puts significant effort into identifying and blocking these spam accounts. Indeed, a recent study of suspended accounts on Twitter shows that Twitter is fairly successful in blocking almost 92% of the spam accounts within 3 days of the first tweet and all of the spam accounts (including those belonging to big spam campaigns) within 6 months [41]. However, to be on the safe side, we do eliminate accounts that have some resemblance to spam account behavior, as reported in [41] (such as followers-to-friends ratio being less than 0.1).

3. CLASSIFYING USERS

In this study, we rely on human knowledge to classify user accounts as Anonymous and Identifiable. In particular, we leverage Amazon Mechanical Turk (AMT). For each Twitter account, we present the account *name* and *screen name* to Mechanical Turk workers and ask them to determine whether these two fields collectively contain (a) just a first name, (b) just a last name, (c) both a first name and a last name, or (d) neither a first nor a last name. The worker can also indicate (e) not sure. We instructed the Mechanical Turk workers to choose 'neither a first nor a last name' and 'both a first name and a last name' options only when they are completely confident, to avoid mis-labelling in situations when there is a lack of clarity (for example due to unusual international names). This enables us to have high confidence in the accounts labelled as not containing names and those containing complete (both first and last) names. To account for human error, we have each account labelled by two Mechanical Turk master workers (those with high ratings). When there is a disagreement, we ask a third master worker to assign the label and use the majority. If there is still a tie among the labels, we (the authors) manually look into the disagreements and finalize the label for the account.

Using these AMT labelings, we define each user account in our data sets as follows:

- **Anonymous** – A Twitter account containing neither the first nor last name (as labelled by AMT) and not containing a URL in the profile (which may point to a web page that identifies or partially identifies the user).
- **Identifiable** – A Twitter account containing both a first name and a last name (as labelled by AMT).
- **Highly Identifiable** – A Twitter account that is Identifiable *and* contains a URL reference to another social network account employing a Real-Name policy (such as Facebook or Google+). It is a subset of the Identifiable group.
- **Partially Anonymous** – A Twitter account having a first name or last name but not both (as labelled by AMT).
- **Unclassifiable** – A Twitter account that is neither Anonymous, Identifiable nor Partially Anonymous. Accounts which have neither a first nor last name but have a URL fall under this category. Also, Twitter accounts that belong to an organization or a company belong here.

We recognize that pseudonymity is different from anonymity, and that Twitter does not support complete anonymity (where the messages are not associated with any pseudonym). However, we prefer to use the more commonly employed term *Anonymous* rather than the more obscure term *Pseudonymous*.

Drawbacks

We mention here that a small fraction of the accounts labelled Anonymous may not be fully anonymous, in that they may provide an identifiable profile photo. However, it has been shown that Twitter profile pictures are often misleading, making it hard to even deduce ethnicity or gender, and are often virtual characters (such as cartoons) or belong to celebrities [37]. Also, a small fraction of the Anonymous users may provide their real identities in their tweets. Furthermore, some users may use fake first and last names, so that a fraction of Identifiable users are effectively Anonymous users. Thus there is some noise in the user classification, noise which is difficult to completely remove. Our results will show, however, that even in the presence of this noise, the Anonymous and Identifiable groups have distinctly different behaviors.

We also point out that employing Amazon Mechanical Turk for user classification is costly in both money and time. (Even if we charge as low as one cent for each account classification, getting multiple workers to label every account adds up for a large-scale study). This limits the number of accounts we can classify, forcing us to optimize our efforts. We are currently exploring techniques for automatic account classification.

4. DATASET COLLECTION AND CHARACTERISTICS

We make use of two distinct data sets in our study.

4.1 Random Accounts

For measuring the prevalence of anonymity in Twitter, we make use of a recent public Twitter dataset released in 2010 containing 41.7 million Twitter accounts [28]. Of the 41.7 million accounts we randomly pick 100,000 accounts and use them as the dataset for this study. It is to be noted that the 2010 public dataset is only used for picking a random subset of Twitter usernames; we use the Twitter API to gather the latest profile information and the friends and follower lists for each of these 100,000 accounts.

We preprocess our initial list of 100,000 users by eliminating all the deactivated accounts, non-English accounts (which do not report English as the language of preference), spam accounts, and ephemeral accounts. The statistics are shown in Table 1. The remaining 50,173 Twitter accounts are passed on to Mechanical Turk for labelling.

4.2 Followers of Sensitive and Non-Sensitive Accounts

We evaluate whether content sensitivity has any correlation with users choosing to be anonymous, by classifying the followers of sensitive and non-sensitive Twitter accounts as

Table 1: Dataset for Measuring Anonymity

Category	# of Twitter Accounts
Deactivated	864
Non-English	5,113
Ephemeral	42,515
Spam	1,335
Remaining	50,173
Total	100,000

Table 3: Labelled Data for Quantifying Anonymity

Label	# of Twitter Accounts
Highly Identifiable	906 (1.8%)
Identifiable	34,085 (67.9%)
Partially Anonymous	10,019 (20%)
Anonymous	2,934 (5.9%)
Unclassifiable	3,135 (6.2%)
Total	50,173

Anonymous and Identifiable. As pointed out in [36], there is no universal definition of what constitutes sensitive content. For this analysis, we create a second dataset by selecting several broad topic categories that are widely considered sensitive and/or controversial by many–pornography, escort services, sexual orientation, religious and racial hatred, online drugs, and guns. We also consider several generic non-sensitive broad categories–news sites, family recreation, movies/theater, kids/babies, and companies/organizations producing household items. For each of these broad categories we identify a few distinctive search terms, and manually pick Twitter accounts that show up when we search for the chosen terms on the Twitter page. When selecting specific accounts in the sensitive categories, we manually look into the account activity to ensure they have high levels of sensitive or controversial tweets.

Most of our short-listed highly-sensitive accounts turned out to have relatively few followers. Among these short-listed accounts, we selected accounts that had at least 200 followers. In total, we picked 50 Twitter accounts related to the different sensitive categories, and 20 accounts related to non-sensitive categories. (Fewer accounts related to nonsensitive categories were needed since those accounts typically have many more followers.) The entire list of chosen Twitter *screen names* in each category and their follower counts are provided in Table 2. Similar to the earlier data collection, to reduce noise we eliminate all non-English, spam and ephemeral followers of these accounts. Because most of the non-sensitive accounts had millions of followers, we conducted our analysis on 1,000 randomly-chosen followers for each Twitter account in the non-sensitive category (to reduce Mechanical Turk costs). All the non-ephemeral followers are again categorized as Identifiable, Partially Anonymous, Anonymous and Unclassifiable using AMT. When comparing different categories, we focus on percentages to ensure that the different numbers of followers do not skew the results.

5. EXPERIMENTAL RESULTS

In this section we report and interpret the results of our experiments.

5.1 Quantifying Anonymity

From our first data set, all the 50,173 accounts (remaining after pre-processing the randomly selected 100,000 Twitter accounts) were labelled using AMT and then categorized as described in Section 3. The distribution of Twitter users across each category is shown in Table 3.

Among the total 50,173 active accounts, we find 5.9% of the accounts are Anonymous. It is to be noted that some of the Identifiable users may contain fake user names and hence actually be anonymous. *Thus, we conclude that anonymity is an important feature for many Twitter users, with at least 5.9% of Twitter users using non-identifiable pseudonyms.* Furthermore, over 25% of the users are semi-anonymous in that they do not provide both their first and last names. *This signifies that online anonymity is important in Twitter, and not having a Real-Name policy could be a strong selling point for a social network.*

The Identifiable user group has 67.9% of the accounts, although as just mentioned, an unknown fraction of these users may actually be anonymous. The Highly Identifiable users, who provide first and last names and link to other social networks with Real-Name policy, constitute 1.8% of the accounts. Although the Highly Identifiable users make up only a small percentage of the Twitter users, we will see they exhibit interesting behavior.

5.1.1 Interests Overlap Between Labelled Groups

To measure whether accounts exhibit similar interests compared to other accounts within the same group, we analyzed the popular friends in the Anonymous and Identifiable categories. We split the Identifiable group into two subsets and compare the friends overlap between the two Identifiable groups, and between the Identifiable and Anonymous groups. Since the Identifiable group is larger, in order to not skew the results, we randomly pick two Identifiable group subsets containing the same number of accounts as the Anonymous group.

Let A denote the set of friends for the accounts in the Anonymous group, and I_1 and I_2 denote the set of friends for the accounts in the two Identifiable groups. For each of the sets of friends, we rank order the friends by popularity. In particular, for each friend $f \in A$, we determine the number of accounts that have f as a friend, and then rank these friends from highest value to lowest value. In an analogous manner, we rank the friends in I_1 and I_2. Then for every top-ranked N friends in each of the three sets, where N varies between 20 and 1000, we determine the overlap. The results are shown in Table 4, where we report the fraction of overlap between the different lists.

Table 4 shows that although there is significant overlap among the popular friends in the Anonymous and the Identifiable groups, for all values of N, the overlap between the Identifiable subsets is always greater. This clearly shows that Anonymous users' interests often deviate from those of Identifiable users. We explore this issue in greater depth in the next subsection.

5.2 Anonymity in Sensitive Accounts

As described in Section 4.2, our second data set consists of 70 accounts (20 non-sensitive and 50 sensitive) along with

Table 2: Sensitive and Non-sensitive Twitter Accounts

Label	Category	Total Followers	Active Followers	Twitter Accounts
Sensitive	Gay/Lesbian	27,315	17,022	GayFollowBack_, blahblah1113, GayDatingFree, GayFlirt, GDates, LorenzoDavids2, GayJockStuds, LoveNudeSelfies, Monstrous10, FreshSX
	Escort Services	11,977	7,113	Escort_Dubai, bocaratonescort, newarkescorts, 001Escort, NYEscorts_Posh, sexinleeds, SapphireEscort, theEscortWeb, glamourescortz, TheEroticGroup
	Pornography	40,261	18,722	bustybethx, MyGayXXXPorn, _youwannafuck, gal_nawty, essexbukkakepar, tattianax, NaughtyTerror, Eritoporn, PeekShowsModels, mysexywifeXXX
	Antisemitism	828	597	againstzionism, We_Hate_Israel
	White Supremacy	3,903	2,218	NiggerHanger, kkkofficial, KKKlan
	Islamophobia	13,834	12,081	banquran2, MuhammadThePig, barenakedislam, KafirCrusaders, IslamExposer
	Marijuana	14,195	11,786	buy_marijuana, BhangChocolate, growweedeasy
	Online Drugs	1,383	1,103	BuyGenericDrugs, buyviagranow, securerxpills
	Guns	8,602	6,835	MyGunsForSale, GunBroker, FirearmsforSale
	Antichristian	1,921	1,292	PriestsRapeBoys
Non-Sensitive	Movies/ Theater	4,000	2,656	aladdin, TheLionKing, DespicableMe, StarTrekMovie
	Family Recreation	4,000	2,933	FamilyFun, FamilyDotCom, NatlParkService, SixFlags
	Companies/ Organizations	4,000	2,242	World_Wildlife, Nestle, LAYS, AOL
	News	4,000	2,634	ReutersLive, abcnews, HuffPostTech, intlCES
	Kids/Babies	4,000	2,929	BabyZone, BabiesRUs, Creativity4Kids, PBSKIDS

Table 4: Popular Friends Overlap Between Anonymous and Identifiable Groups

# of Top Popular Friends (N)	Fraction of Overlap		
	$\frac{I_1 \cap I_2}{N}$	$\frac{A \cap I_1}{N}$	$\frac{A \cap I_2}{N}$
20	0.9	0.55	0.55
30	0.93	0.57	0.57
50	0.88	0.62	0.64
70	0.87	0.66	0.66
100	0.84	0.65	0.64
200	0.87	0.68	0.71
500	0.87	0.69	0.71
1000	0.84	0.66	0.69

all the followers of these accounts, as summarized in Table 2. Leveraging AMT, each follower is categorized as Anonymous, Partially Anonymous, Identifiable, or Unclassifiable. Figure 1 shows the average percentage of followers who are Anonymous, Identifiable and Highly Identifiable (subset of Identifiable) for each category of sensitive and non-sensitive accounts. The categories are arranged in order from the highest percentage to the lowest percentage of Anonymous followers.

We first observe that the different categories contain greatly different percentages of Anonymous and Identifiable followers. The percentage of Anonymous users varies from 6.6% to 37.3%; the percentage of Identifiable users varies from 26.9% to 59.6%. Strikingly, the sensitive categories have the largest percentage of Anonymous users. Except for Online Drugs, all of the sensitive categories have more than 10.3% of Anonymous followers and all the non-sensitive categories have at most 8.9% of Anonymous followers. Pornography, Marijuana, Islamophobia and Gay/Lesbian all have more than 21.6% of Anonymous followers, with pornography far exceeding the rest with 37.3% of Anonymous followers.

For the percentages of Identifiable followers, there are also patterns, although not as clearly demarcated as for the Anonymous percentages. The categories with fewer than 40% of Identifiable followers (Pornography, Marijuana, Gay/ Lesbian, Escort Groups) are all sensitive; and most of the categories with more than 50% of Identifiable followers are non-sensitive categories. But some of the sensitive categories have a surprisingly large percentage of Identifiable followers (e.g. White Supremacy and Guns). We believe one reason the patterns may be less strong for Identifiable users is because the Identifiable category may be noisier than the Anonymous category, as a significant fraction of the Identifiable users may be using fake names and are in actuality Anonymous. It is also possible that many followers in the White Supremacy and Guns categories take "pride" in being

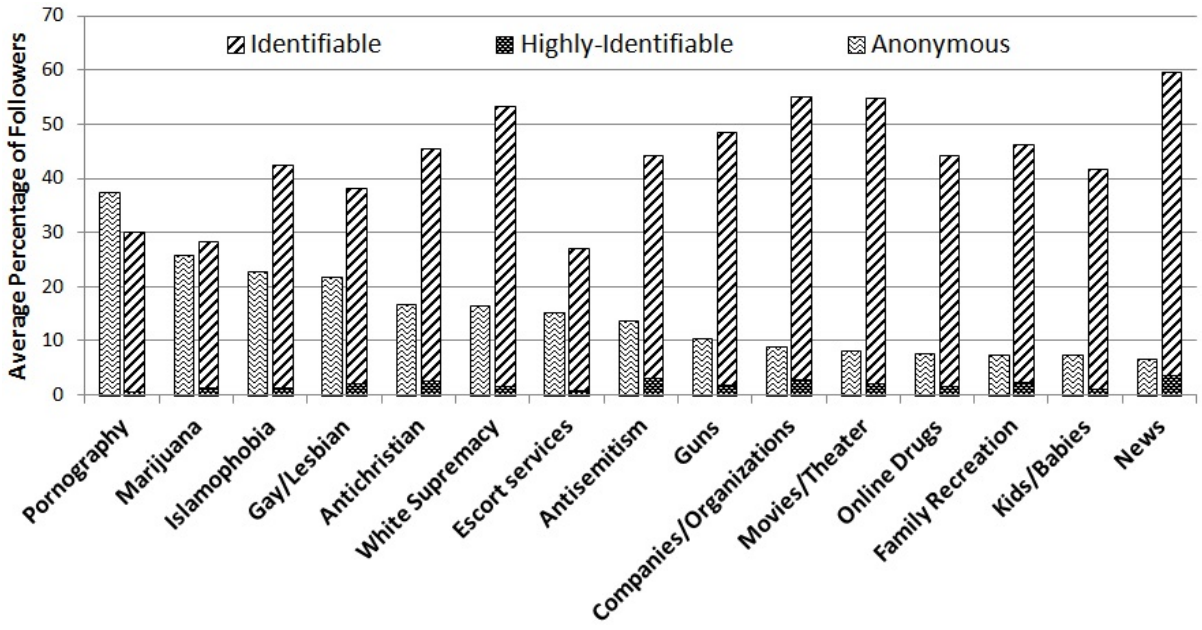


Figure 1: Sensitive and Non-Sensitive Twitter Account Categories: Follower Distribution

members of these groups and do not feel the need to hide their identities. This shows that there are different types of sensitive content–while some generate secrecy, others may influence people to be open (resulting in having many Identifiable followers rather than Anonymous). This establishes that content sensitivity is quite nuanced and complex. For the Highly Identifiable users, there is less of a pattern, although Pornography and Escort Services have the lowest percentages.

Based on these patterns, we can define a simple sensitive *topic* classifier that relies on the percentage of Anonymous and Identifiable followers (for example, (i) a topic with more than 10% of Anonymous followers is potentially sensitive or controversial, and (ii) a topic with more than 40% of Identifiable followers and fewer than 10% of Anonymous followers is likely non-sensitive). In future work, we expect to build an automated mechanism to determine whether an account is Anonymous, Identifiable, or Unclassifiable, which can then be used for topic sensitivity classification.

Even non-sensitive categories have 6.6% – 8.9% of Anonymous followers. This is an important observation that validates that users do not create anonymous profiles for the *sole* purpose of following sensitive accounts. To avoid maintaining multiple profiles, an Anonymous user might follow both sensitive and non-sensitive Twitter accounts using the same profile, leaking out his interests on Twitter. For example, by following the *Star Wars Movie* Twitter account, a user indicates that he is interested in the *Star Wars* franchise. These non-sensitive interest disclosures can potentially be used to deanonymize the Anonymous users using techniques similar to [45] (which shows that a user's group membership sets in a social network are generally unique and can be used to identify the user).

Figure 1 presents results for the accounts aggregated over each category. To gain further insight, we now consider the *individual* accounts instead of the account categories. We consider all of the non-sensitive accounts and all of the highly-sensitive accounts, except those belonging to the online drugs category (which appears from the data to not actually be a highly-sensitive category). Figure 2 shows a scatterplot with one point for each of these accounts. The x-axis of the plot indicates the fraction of Identifiable followers, the y-axis indicates the fraction of Anonymous followers. From the figure we can see that there is a general trend to have more Anonymous followers for the sensitive accounts and more Identifiable followers for the non-sensitive accounts. In fact, we see that the line $y = 0.905 \times x - 0.305$ (obtained using linear regression) separates the points belonging to the two classes–all but three of the sensitive accounts are above the line, and all but one of the non-sensitive accounts are below this line. In future work, we expect to build an automated mechanism to determine whether an account is Anonymous, Identifiable, or Unclassifiable. Figure 2 provides hope that we may be able to develop a sensitive account detector based on the percentages of detected Anonymous and Identifiable followers.

Twitter's popularity has resulted in an increase in its misuse. With the help of legal authorities, Twitter management is actively fighting spam [17, 41], spread of pirated media content [26], child porn [1], and terrorism [44]. As most of the miscreants already employ evasion techniques against current detection mechanisms–such as keyword based detection or URL spam/phishing detection–it becomes important to identify new signals that can be leveraged. Figures 1 and 2 show there is indeed a strong correlation between follower anonymity and the account sensitivity. This validates that analyzing Anonymous followers can help us detect these sensitive accounts–helping narrow down the illegal and controversial account search space. This approach does not replace existing detection techniques, but is complementary and helps raise the bar for miscreants.

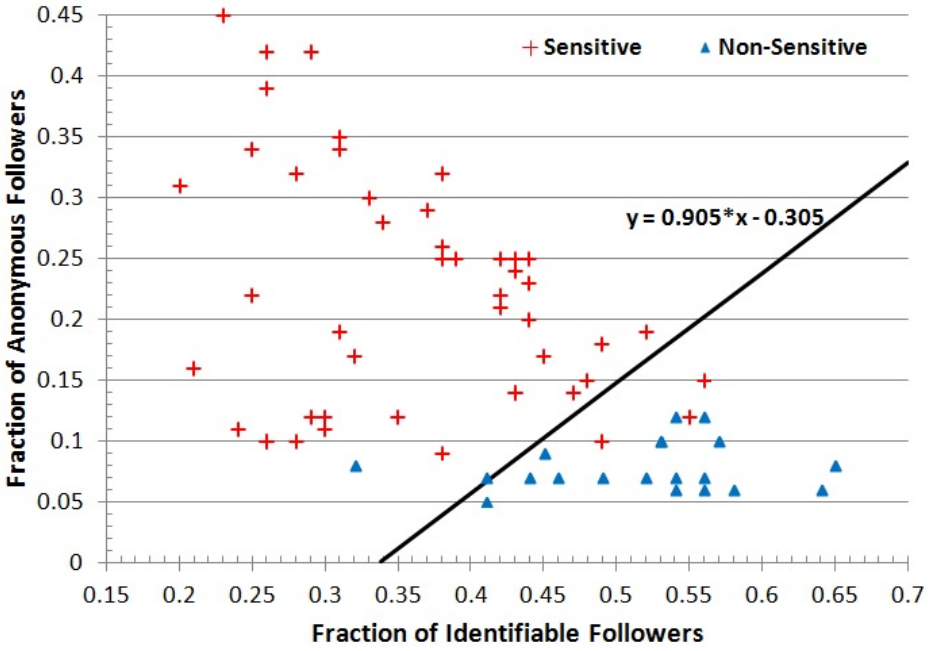


Figure 2: Sensitive and Non-Sensitive Twitter Accounts: Scatter Plot

Table 3 indicates that the percentage of Anonymous users is 5.9%, whereas Figure 1 shows that even non-sensitive accounts have Anonymous users going up to 8.9%. Similarly, the percentage of Identifiable users in Table 3 is 68%, where as in the second dataset they do not go beyond 60%. One reason for these small differences could be the difference in the age of the accounts in the two datasets. The second dataset has many recently created accounts (median account creation date is *Jan 09, 2011*) compared to the first dataset (median is *Apr 22, 2009*).

5.3 Behavioral Analysis

For the behavioral analysis of Twitter accounts we combine the datasets from the earlier two studies. After eliminating all the non-English, spam and ephemeral accounts, the distribution of labelled accounts across each category is shown in Table 5.

5.3.1 Lurker and Protected Accounts

Many OSNs have silent participants. We categorize a Twitter user as *Lurker* if the user does not post any tweets. Since ephemeral users have been removed in this study, a Lurker therefore is a user who has been active for at least six months (as evidenced by adding a friend at least six months after account creation) but yet has never posted a tweet.

As stated in Section 2, Twitter supports a *protected* privacy feature that enables users to protect their activity from being publicly visible. We investigate whether Anonymous users make use of this *protected* feature. Table 5 shows the distribution of Protected and Lurker accounts across the different labelled groups (Unclassifiable and Partially Anonymous are not shown).

We see from Table 5 that Identifiable users have a greater tendency to be private as compared to Anonymous users. A reasonable explanation for this is that, because an Identifiable user makes his identity known through his profile, he may be more reluctant to publicly share his tweets and friend list with the public at large. We also see that Identifiable users have a greater tendency to lurk as compared to Anonymous users. An Anonymous user may be less inhibited about tweeting, and hence is more willing to tweet than an Identifiable user.

Table 5: Protected and Lurker Twitter Account Statistics

Label	# of Twitter Accounts	Protected Accounts	Lurker Accounts
Highly Identifiable	2,082	301 (14.5%)	7 (0.3%)
Identifiable	65,293	8,547 (13.1%)	2,895 (4.4%)
Anonymous	19,942	2,035 (10.2%)	592 (3%)

Combining the datasets in Section 5.1 and Section 5.2 (containing many sensitive accounts), the Anonymous users increase from 5.9% to 15.3%. The Identifiable users decrease from 67.9% to 50%. This re-emphasizes that content sensitivity has a strong correlation with user anonymity. However, currently we do not know which is the cause and which is the effect, i.e., do users create Anonymous accounts because they want to follow sensitive topics (we identified that it is not the *sole* reason earlier in Section 5.2), or is it because they have an Anonymous account they do not shy away and are more open in expressing sensitive interests. This causality issue remains an open question.

5.3.2 Friends, Followers, and Tweets

To measure how different labelled groups use Twitter product features, we look at the friend, follower and tweet statistics, shown in Figure 3.

From the graphs, we can see that Highly Identifiable users have more friends, more followers and even post more tweets –implying that they are very socially active. The Identifiable users are at the other extreme–they have fewer friends, followers, and tweets. The Anonymous users have some similarities with the Highly Identifiable users, but are distinct compared to the Identifiable users.

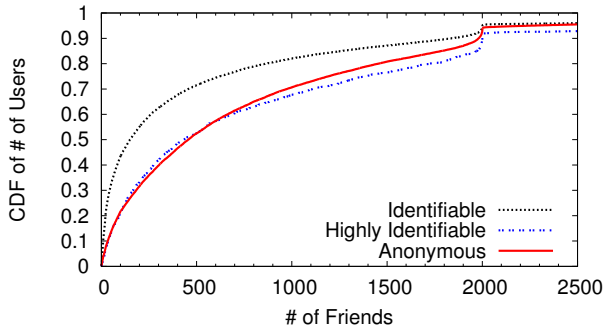
From Figure 3a, we can see that Anonymous users tend to have many more friends (i.e., follow many more people) than Identifiable users. Being unidentifiable allows them to follow many accounts, including sensitive accounts, without worrying about the repercussions. The Identifiable users are perhaps more conservative in choosing who to follow, as it is possible to trace back their online actions to their real world identities. The kink in the CDFs at 2000 is due to Twitter's famous *follow limit*[1]–preventing people from following more than 2000 accounts, only exceedable by having many followers.

Figure 3b also shows that being Anonymous does not negatively impact the online user experience, as users are still able to obtain many followers. As Anonymous users do not hold back when sharing information or expressing opinions online, they seem to be much better at building an "online brand" for themselves, thereby attracting more users to follow them. Figure 3c indicates that Anonymous users post more tweets than Identifiable users, but they are not as socially active as Highly Identifiable users.

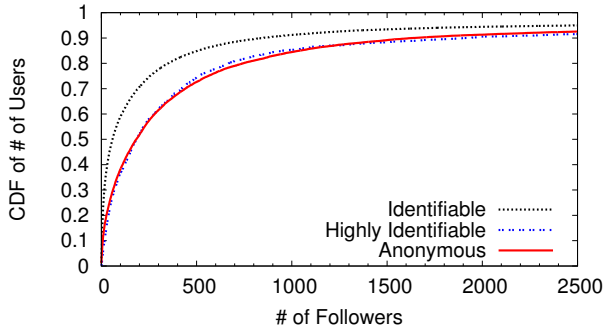
The median number of friends for Highly Identifiable and Anonymous groups are at 432 and 456.6, whereas the Identifiable group is far behind at 151. Similarly, the median number of followers of Highly Identifiable and Anonymous groups are very close–184.5 and 184. The median number of followers for Identifiable group is just 59. In the number of tweets, we see some distinction between Anonymous and Highly Identifiable groups–the medians of Identifiable, Anonymous and Highly Identifiable are at 145, 423 and 790 respectively.

Table 6 shows the friends and followers statistics for the Lurkers belonging to different labelled categories. Anonymous Lurkers are very active compared to Lurkers in Identifiable group–on average, they have nearly double the number of friends and followers. The strange behavior of many people following an Anonymous account, which does not post any tweets, is likely due to the attractiveness of the profile information and the expectation that something interesting (or sensitive) might be posted by that account. The same excitement does not hold for an Identifiable account.

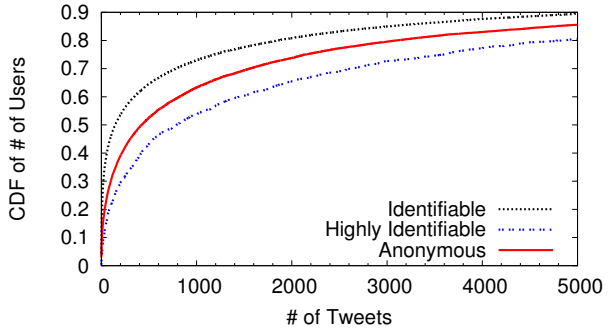
The main takeaway message in this subsection is that Anonymous users are generally more active participants than Identifiable users, as they tweet more, lurk less, follow more accounts, and are more willing to expose their activity to the general public. These findings indicate that Anonymous and Identifiable users exhibit different online behaviors, and shows the feasibility of using these behaviors in developing an automatic Anonymous and Identifiable account classifier.



(a) Friends Distribution



(b) Followers Distribution



(c) Tweets Distribution

Figure 3: Friends, Followers and Tweets Statistics

6. FUTURE DIRECTIONS

As mentioned in Section 3, relying on AMT significantly limits the number of accounts we can analyze. Having a larger sample of labelled accounts can help us (i) automatically detect sensitive accounts (such as those spreading child porn–as outlined in Section 5.2), (ii) better understand the nuances of content sensitivity and its influence on online user behavior, (iii) better understand the reasons for choosing anonymous pseudonyms or evaluate the after effects of the choice, or (iv) identify behavioral traits that can be used to deanonymize the Anonymous Twitter users. Automatic user classification–possible through building large first and last name lists (e.g., crawling Facebook/Google+ user directories), or building efficient machine learning classifiers after studying a small ground truth labelled set–can help us overcome the limitations. We expect to pursue this direction in the future.

Tweets are a very important source of information that we did not exploit in this work. A user may reveal many of the user's private attributes–such as name, gender, age, sexual preference, etc.–in the tweets. Incorporating tweets into our anonymity study can help reduce the noise in the dataset. Furthermore, studying tweets can help evaluate whether anonymity has any correlation with making controversial posts on Twitter.

[1]`https://support.twitter.com/articles/66885-why-can-t-i-follow-people`

Table 6: Friends and Followers Statistics for Different Lurker Categories

	# of Friends			# of Followers		
Label	**Mean**	**Median**	**Standard Deviation**	**Mean**	**Median**	**Standard Deviation**
Identifiable Lurkers	64.8	20	217.2	15.9	5	76.8
Anonymous Lurkers	182.6	39	437.8	57.4	5	361.9

7. RELATED WORK

7.1 Surveys and Interviews

Employing surveys and interviews, many social scientists, economists, psychologists have taken interest in online anonymity [29, 42, 11, 23, 39, 10]. Our study differs from these approaches in that we employ a large-scale data-driven approach, rather than rely on surveys and interviews. The large-scale approach not only allows us to quantify issues in anonymity and reduce statistical errors, but also permits us to explore new issues, such as the importance of following sensitive topics as a motivation for anonymity. Moreover, as argued in the paper, the data-driven approach may allow us to automatically identify sensitive topics and controversial accounts (and perhaps illegal activity).

In particular, a recent survey-based study points out that people are actively seeking anonymity on the web, and that it increases online engagement [24]. This result mirrors our data driven results in Section 5.2, where we show that Anonymous users are generally less inhibited to be active participants, as they tweet more, lurk less, follow more accounts, and are more willing to expose their activity to the general public. Acquisti et al. employ an analytical framework to investigate the economics of anonymity, so it can be introduced as a feature in many online applications [4].

7.2 Data Driven Studies

There are several recent data-driven studies of anonymity using blogs and web sites. Gomez et al. studied anonymous user comments on the technology news website, Slashdot, and found that fully anonymous comments made up 18.6% of the total, and that pseudonymity is the norm when reputation mechanisms are enforced [16]. A study about 4chan, an image board website, showed that online communities can succeed despite being fully anonymous and extremely ephemeral [5]. By analyzing a question-and-answer website–Quora, Peddinti et al. show that it is possible to gain a novel understanding of users' perspectives on content sensitivity via a data-driven analysis of their usage of anonymity features of the website [36].

In our work, we focus on anonymity in online social networks. We measure the extent to which people exercise anonymity when provided with a choice of being identifiable or anonymous, as well as evaluate the behavior of anonymous and identifiable users. To our knowledge, we are the first to perform a data-driven analysis of user anonymity in online social networks.

7.3 Anonymization and Deanonymization

While some researchers have been trying to improve the anonymity guarantees in a social network [40, 7], there is a large body of research work on deanonymizing users and linking users across different social networks. Some deanonymization techniques link different social network accounts using usernames [38, 46], whereas others rely on social connection graphs [35].

Techniques such as [19] and [43], compare profile information across multiple websites to identify and link accounts belonging to the same user, thereby generating a richer user profile than is possible to infer from a single website. [15] and [20] make use of the user posted content on the website to link different social network profiles. Solutions involving machine learning techniques have also been proposed to disambiguate profiles belonging to the same user across different social networks [32, 30, 27].

Some of these techniques can potentially be used to deanonymize anonymous users on Twitter. However, the focus of our work has not been on deanonymizing users, but rather on quantifying people's desire to be anonymous and to study the behavior of anonymous users.

7.4 Studies on Twitter

Due to the availability of an API, many have extensively studied different aspects of Twitter–Java et al. focus on how Twitter users choose others to follow [21], Kwak et al. study information dissemination [28], Cha et al. measure user influence [9], and Castillo et al. study information credibility [8]. Studies have also been undertaken to improve usability of Twitter, such as developing a user recommendation service [18]. We study how Twitter users utilize the in-built privacy features.

8. CONCLUSION

We performed a large-scale analysis of Twitter to study the prevalence and behavior of Anonymous and Identifiable users. We employed Amazon Mechanical Turk (AMT) to classify Twitter users as Highly Identifiable, Identifiable, Partially Anonymous, and Anonymous. We quantified the importance of Anonymity for a large fraction of online users. We then selected several broad topic categories that are widely considered sensitive and found that there is a correlation between content sensitivity and a user's choice to be anonymous. Finally, we found that Anonymous users are generally less inhibited to be active participants, as they tweet more, lurk less, follow more accounts, and are more willing to expose their activity to the general public.

9. ACKNOWLEDGEMENTS

We thank Aleksandra Korolova for valuable feedback on the paper draft.

This work was supported in part by the NSF (under grant CNS-1318659). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any of the sponsors.

10. REFERENCES

[1] Child pornography via Tweet: pedophiles abuse Twitter as a distribution channel. http://www.naiin.org/en/news/Child-pornography-via-Tweet-pedophiles-abuse-Twitter-as-a-distribution-channel-75.html.

[2] Create your Google+ profile name. https://support.google.com/plus/answer/1228271?hl=en. Accessed: Feb 8th, 2014.

[3] Facebook's Name Policy. https://www.facebook.com/help/292517374180078. Accessed: Feb 8th, 2014.

[4] A. Acquisti, R. Dingledine, and P. Syverson. On the economics of anonymity. In *Financial Cryptography*, Lecture Notes in Computer Science, 2003.

[5] M. S. Bernstein, A. Monroy-Hernández, D. Harry, P. André, K. Panovich, and G. G. Vargas. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[6] V. Blue. Forced Google Plus integration on YouTube backfires, petition hits 112,000. http://www.zdnet.com/forced-google-plus-integration-on-youtube-backfires-petition-hits-112000-7000023196/.

[7] A. Campan and T. Truta. Data and structural k-anonymity in social networks. In *Proceedings of Privacy, Security, and Trust in KDD*, 2009.

[8] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, 2011.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[10] T. Chesney and D. K. Su. The impact of anonymity on weblog credibility. *International Journal of Human-Computer Studies*, 68(10), 2010.

[11] T. Connolly, L. M. Jessup, and J. S. Valacich. Effects of anonymity and evaluative tone on idea generation in computer-mediated groups. *Management Science*, 36(6), 1990.

[12] B. Dowell. Rise in defamation cases involving blogs and Twitter. http://www.theguardian.com/media/2011/aug/26/defamation-cases-twitter-blogs.

[13] C. GAYLORD. Facebook's Forgotten Rule: No Fake Names Allowed. http://abcnews.go.com/Technology/facebooks-forgotten-rule-fake-names-allowed/story?id=15509496.

[14] D. GEERE. Twitter spread misinformation faster than truth in UK riots. http://www.wired.co.uk/news/archive/2011-08/09/twitter-misinformation-riots.

[15] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, 2013.

[16] V. Gómez, A. Kaltenbrunner, and V. López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th International Conference on World Wide Web (WWW)*, 2008.

[17] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS)*, 2010.

[18] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. Wtf: The who to follow service at twitter. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, 2013.

[19] D. Irani, S. Webb, K. Li, and C. Pu. Large online social footprints–an emerging threat. In *Proceedings of International Conference on Computational Science and Engineering (CSE)*, 2009.

[20] P. Jain, P. Kumaraguru, and A. Joshi. @I Seek 'Fb.Me': Identifying Users Across Multiple Online Social Networks. In *Proceedings of the 22nd International Conference on World Wide Web Companion (WWW Companion)*, 2013.

[21] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*, 2007.

[22] A. Jeffries. Facebook's fake-name fight grows as users skirt the rules. http://www.theverge.com/2012/9/17/3322436/facebook-fake-name-pseudonym-middle-name.

[23] L. M. Jessup, T. Connolly, and J. Galegher. The effects of anonymity on gdss group process with an idea-generating task. *MIS Q.*, 14(3), 1990.

[24] R. Kang, S. Brown, and S. Kiesler. Why do people seek anonymity on the internet?: informing policy and design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2013.

[25] A. Kavanaugh, S. Yang, S. Sheetz, L. T. Li, and E. Fox. Microblogging in crisis situations: Mass protests in Iran, Tunisia, Egypt. In *Workshop on Transnational Human-Computer Interaction, CHI*, 2011.

[26] S. Knafo and J. Bialer. How Twitter Handles Piracy – An Inside Look. http://www.huffingtonpost.com/2012/02/02/how-twitter-handles-piracy_n_1251167.html.

[27] M. Korayem and D. J. Crandall. De-anonymizing users across heterogeneous social computing platforms. In *Proceedings of 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013.

[28] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, 2010.

[29] Y. Lelkes, J. A. Krosnick, D. M. Marx, C. M. Judd, and B. Park. Complete anonymity compromises the accuracy of self-reports. *Journal of Experimental Social Psychology*, 48(6), 2012.

[30] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon. What's in a name?: an unsupervised approach to link users across communities. In

Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM), 2013.

[31] N. Lomas. Facebook Users Must Be Allowed To Use Pseudonyms, Says German Privacy Regulator; Real-Name Policy 'Erodes Online Freedoms'. `http://techcrunch.com/2012/12/18/facebook-users-must-be-allowed-to-use-pseudonyms-says-german-privacy-regulator-real-name-policy-erodes-online-freedoms/`.

[32] A. Malhotra, L. C. Totti, W. M. Jr., P. Kumaraguru, and V. Almeida. Studying user footprints in different online social networks. *CoRR*, abs/1301.6870, 2013.

[33] E. Munoz. NYPD Twitter campaign implodes, flooded with photos of police abuse. `http://rt.com/usa/154120-nypd-hashtag-twitter-police/`.

[34] E. Mustafaraj, P. T. Metaxas, S. Finn, and A. Monroy-HernÃąndez. Hiding in plain sight: A tale of trust and mistrust inside a community of citizen reporters. In *Proceedings of 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[35] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Proceedings of 30th IEEE Symposium on Security & Privacy*, 2009.

[36] S. T. Peddinti, A. Korolova, E. Bursztein, and G. Sampemane. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In *Proceedings of the 35th IEEE Symposium on Security & Privacy*, 2014.

[37] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[38] D. Perito, C. Castelluccia, M. Kaafar, and P. Manils. How unique and traceable are usernames? In *Proceedings of the 11th International Conference on Privacy Enhancing Technologies (PETS)*. 2011.

[39] T. Postmes, R. Spears, K. Sakhel, and D. de Groot. Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, 27(10), 2001.

[40] K. P. Puttaswamy, A. Sala, and B. Y. Zhao. Starclique: Guaranteeing user privacy in social networks against intersection attacks. In *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, 2009.

[41] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC)*, 2011.

[42] P. A. Thompsen and D.-K. Ahn. To be or not to be: An exploration of e-prime, copula deletion and flaming in electronic mail. In *ETC: A Review of General Semantics;Summer92, Vol. 49 Issue 2, p146*, 1992.

[43] J. Vosecky, D. Hong, and V. Shen. User identification across multiple social networks. In *Proceedings of First International Conference on Networked Digital Technologies (NDT)*, 2009.

[44] D. Wiener-Bronner. Twitter Is the Preferred Social Media Platform Among Terrorists. `http://www.businessinsider.com/terror-groups-twitter-2014-5`.

[45] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A practical attack to de-anonymize social network users. In *Proceedings of 31st IEEE Symposium on Security & Privacy*, 2010.

[46] R. Zafarani and H. Liu. Connecting corresponding identities across communities. In *Proceedings of 3rd International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.