The Tweets They are a-Changin': Evolution of Twitter Users and Behavior

Yabing Liu[†] **Chloe Kliman-Silver**[‡] **Alan Mislove**[†]

[†]Northeastern University [‡]Brown University

Abstract

The microblogging site Twitter is now one of the most popular Web destinations. Due to the relative ease of data access, there has been significant research based on Twitter data, ranging from measuring the spread of ideas through society to predicting the behavior of real-world phenomena such as the stock market. Unfortunately, relatively little work has studied the changes in the Twitter ecosystem itself; most research that uses Twitter data is typically based on a small time-window of data, generally ranging from a few weeks to a few months. Twitter is known to have evolved significantly since its founding, and it remains unclear whether prior results still hold, and whether the (often implicit) assumptions of proposed systems are still valid.

In this paper, we take a first step towards answering these question by focusing on the evolution of Twitter's users and their behavior. Using a set of over 37 billion tweets spanning over seven years, we quantify how the users, their behavior, and the site as a whole have evolved. We observe and quantify a number of trends including the spread of Twitter across the globe, the rise of spam and malicious behavior, the rapid adoption of tweeting conventions, and the shift from desktop to mobile usage. Our results can be used to interpret and calibrate previous Twitter work, as well as to make future projections of the site as a whole.

Introduction

Online social networks(OSNs) are now a popular way for users to connect, communicate, and share content; many serve as the de-facto Internet portal for millions of users (Post 2014). Because of the massive popularity of these sites, data about the users and their communication offers unprecedented opportunities to examine how human society functions at scale. As a result, significant recent research has focused on these sites, with a particular emphasis on Twitter due to the relative ease of accessing a large amount of data. For example, recent papers have ranged from studying Twitter-specific behavior (e.g., the patterns of retweeting (Macskassy and Michelson 2011), deletion of tweets (Almuhimedi et al. 2013) and usage of different entities in tweet contents (Macskassy 2012; Yang et al. 2012)), to examining how privacy leaks and violations can occur (Mao, Shuai, and Kapadia 2011; Meeder et al. 2010), and even using aggregated data from Twitter to predict the behavior of real-world phenomena such as the stock market (Gilbert and Karahalios 2010).

While the set of research using Twitter data has expanded rapidly, there has been relatively little work that has studied the evolution of the Twitter ecosystem *itself*. For example, Twitter has grown from thousands of users in 2007 to millions in 2009 to hundreds of millions in 2013. In parallel with this growth, we have seen a significant maturation of the Twitter platform: Twitter today is used by many organizations and individuals as a primary way of communicating with others. Most research that uses Twitter data is typically based on a small time-window of data—generally ranging from a few weeks to a few months—making it difficult to quantify long-term trends. Twitter is known to have evolved significantly since its founding, and it remains unclear how much the user base and behavior has evolved, whether prior results still hold, and whether the (often implicit) assumptions of proposed systems are still valid.

In this paper, we take a first step towards answering these questions by directly examining the evolution of Twitter itself, focusing on the Twitter users and their behavior. Using a set of over 37 billion tweets spanning between 2006 and 2013, we quantify how the users, their behavior, and the site as a whole have evolved. We observe and quantify a number of interesting patterns, including

- The spread of Twitter across the globe, both in terms of users in different regions and tweets containing different languages (e.g., the percentage of U.S./Canada users drops from over 80% to 32%, and the percentage of users tweeting in English falls from 83% to 52%).
- The percentage of tweets that are no longer available due to a user's or Twitter's action increases to over 20% for some time ranges.
- The percentage of Twitter user accounts today that are inactive shows rapid growth; over 32.5% of accounts have not tweeted for over a year.
- The increase of malicious behavior on Twitter beginning in 2009, including fake followers, fake accounts, and hashtag promotion; over 6% of all accounts are now suspended.

Dataset	Date range	Users	Tweets	Date collected	Coverage of all Tweets	Coverage of all Users
Crawl	21/03/2006 – 14/08/2009	25,437,870	1,412,317,185	14/08/2009	∼100%	∼100%
Gardenhose	15/08/2009 – 31/12/2013	376,876,673	36,495,528,785	Time of tweet	∼10–15%	∼30.61%
UserSample	21/03/2006 – 31/12/2013	1,210,077	—	12/31/2013	∼0.1%	∼0.1%
Total	21/03/2006 – 31/12/2013	388,796,600	37,907,845,970	—	—	—

Table 1: Source and basic statistics for the Twitter data used in this study.

- The switch from a primarily-mobile system (via SMS) to a primarily-desktop system (via the web) and back to a primarily-mobile system (via mobile applications). Today, over half of all tweets are created on mobile devices.

Our results can be used to interpret and calibrate previous Twitter studies, as well as to make future projections of the site as a whole. We make all of our analysis available to the research community (to the extent allowed by Twitter's Terms of Service) to aid other researchers and to stimulate further research in this area; researchers can access it at

`http://twitter-research.ccs.neu.edu/`

Background and Data Source

Twitter is a "micro-blogging" service that allows users to multicast short messages (called *tweets*). Each user has a set of other users (called *followers*) who receive their messages; those who a user follows are called *friends*. The follow relationship in Twitter is directed, and requires authorization from the followee only when the followee has elected to make their account *protected*. Each tweet can only be up to 140 characters in length, and the default setting in Twitter is to allow all tweets to be publicly visible.

Twitter data We obtain our Twitter data from two sources; basic statistics of these datasets is provided in Table 1.[1] First, we use an almost-complete[2] collection of all tweets issued between March 21, 2006 and August 14, 2009 collected by previous work (Cha et al. 2010); we refer to this as the *Crawl* dataset. This dataset was collected in August 2009 by iteratively downloading all of the tweets of all public users alive at the time.

Second, we collect data from the Twitter "gardenhose" public stream[3] between August 15, 2009 and December 31, 2013; we refer to this as the *Gardenhose* dataset. Our measurement infrastructure was down for the 10 weeks between October 18, 2010 and December 31, 2010, so we do not have data for that time period. Twitter states that the gardenhose is a random sample of all public tweets. Each *Gardenhose* tweet includes information about the user who created the tweet (e.g., the user's location and count of total tweets) that is current as-of the time of the tweet.

[1] This study was conducted under Northeastern University Institutional Review Board protocol #10-03-26.

[2] The dataset does not include any tweets deleted before August 14, 2009, and only includes the 3,200 most recent tweets (as of August 14, 2009) for each user.

[3] `https://stream.twitter.com/1.1/statuses/sample.json`, with elevated access.

Gardenhose sampling rate We briefly estimate the sampling rate of the *Gardenhose* dataset; Twitter states that the gardenhose is a random sample, but does not state the rate. We estimate the sampling rate by relying on the `statuses_count` field of the user in each tweet; the `statuses_count` field represents the total number of tweets (statuses) that the user has issued at the time of the tweet. Each month, we determine the first observed value of `statuses_count` (sc_{first}), the last observed value of `statuses_count` (sc_{last}), and the number of tweets we observed (obs). We can then estimate the sampling rate for that user with

$$rate = \frac{obs}{sc_{last} - sc_{first}}$$

We plot the average value of $rate$ across all users with $sc_{last} - sc_{first} > 1000$ in Figure 1, observing a sampling rate of ∼15% until 07/2010, and ∼10% since then (the "dips" observed in the graph are due to short periods of time when our measurement infrastructure was down).

Limitations Because the *Crawl* dataset was collected in August 2009 (as opposed to the *Gardenhose* dataset, which was collected over a period of years, as tweets were issued), the user information for the *Crawl* dataset is as-of August 2009. This limitation will occasionally present itself during our analysis, and we discuss these limitations in-line.

We also face two limitations with the *Gardenhose* dataset. First, the *Gardenhose* dataset ends up containing biased sample of users, with a bias towards more active users. The reason for this is that users who tweet very often are extremely likely to show up in our dataset; a user who tweets only once has a ∼10% chance of appearing. Second, Twitter does not inform us when users leave the network, so we are

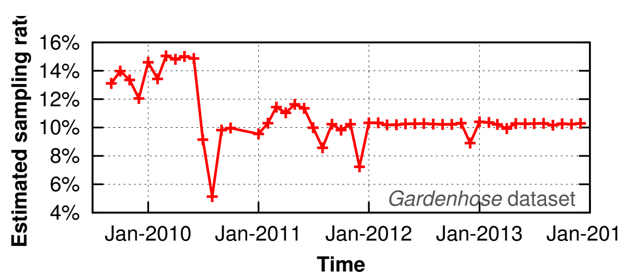

Figure 1: The estimated sampling rate (the average value of *rate* for users with more than 1,000 statues in a month) of the *Gardenhose* dataset over time. The occasional drops in sampling rate are due to times that our collection infrastructure was down for short periods of time.

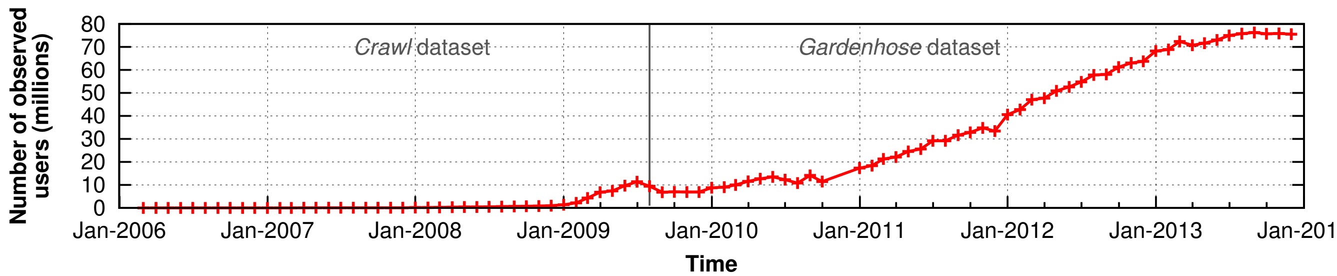

Figure 2: The number of users we observed tweeting in each month in the *Crawl* and *Gardenhose* datasets. The "dip" in August 2009 is due to the switch from a complete sample to a 15% sample. Our numbers are much lower that Twitter's announced "active users" numbers due to our data sample and Twitter's definition of an "active" user.

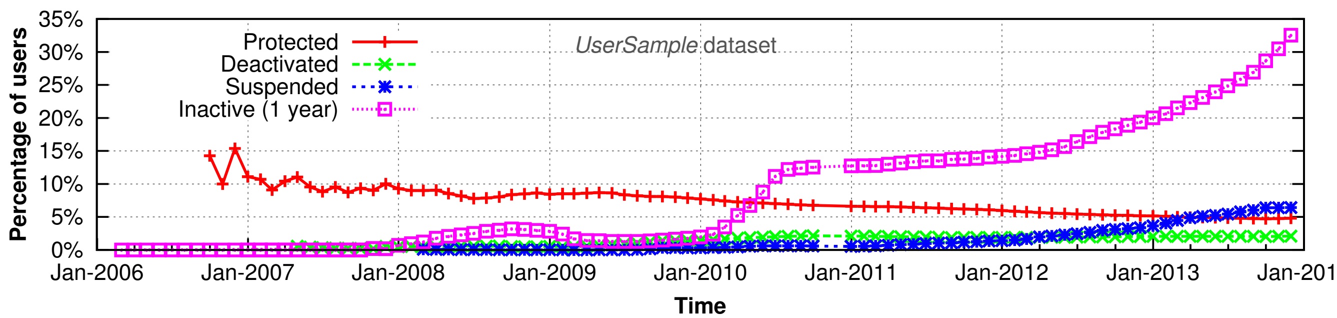

Figure 3: The percentage of the entire Twitter user base over time whose accounts are protected, deactivated, suspended, or inactive (for at least a year), based on the *UserSample* dataset. We observe a dramatic increase in both inactive and suspended accounts starting in 2010.

unable to determine when users mark their accounts as protected (thereby hiding their tweets), are *suspended* by Twitter for violating the Terms of Service, or *deactivate* their accounts (i.e., manually delete their account).

To address these limitations of the *Gardenhose* dataset, we collect a third and final dataset *UserSample* that represents a random sample of *users* instead of *tweets*. Specifically, we generate 2 million random `user_ids` between 1 and 1,918,524,009 (the largest `user_id` that we ever observed), representing a ~0.1% sample of all Twitter users. We then query Twitter (both via the API and the web site) in January 2014 for the most recent information on each of these users, allowing us to determine if the `user_id` was ever assigned, has been suspended, or is protected. We infer if the `user_id` has been deactivated by the user if Twitter says the user does not exist but we observed a tweet from the user in our *Crawl* or *Gardenhose* datasets. Overall, we find that at least 1,210,077 (60.51%) of these 2 million `user_ids` were ever assigned to a user.

Throughout our analysis, we use the most appropriate dataset(s) for each question at hand. Additionally, we label each graph with the dataset(s) that it uses.

User characteristics

We begin our analysis by studying how the user population of Twitter has changed since its inception.

User growth and activity We first examine the characteristics of the Twitter user population. Figure 2 shows the total number of users that we observed over time; we observe massive growth that is in-line with Twitter's reported number of monthly active users (Weil 2010). For example, we observed over 73 million users tweet in June 2013; while Twitter reports 218 million active users for that time period (SEC 2013); our number is lower due to the fact that we have a random 10% sample and Twitter's definition of an active user is based on login activity, not tweeting activity. We also observe rapid growth from 2009 through 2012, with a leveling-off of the number of active users in 2013; this suggests that Twitter's user population growth may be

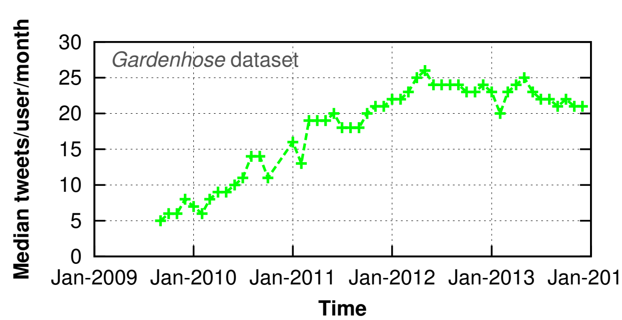

Figure 4: The median number of tweets per user per month over time, based on the first and last `statuses_count` observed for each user. Note that this result is based on *Gardenhose* dataset, which is biased towards more active users.

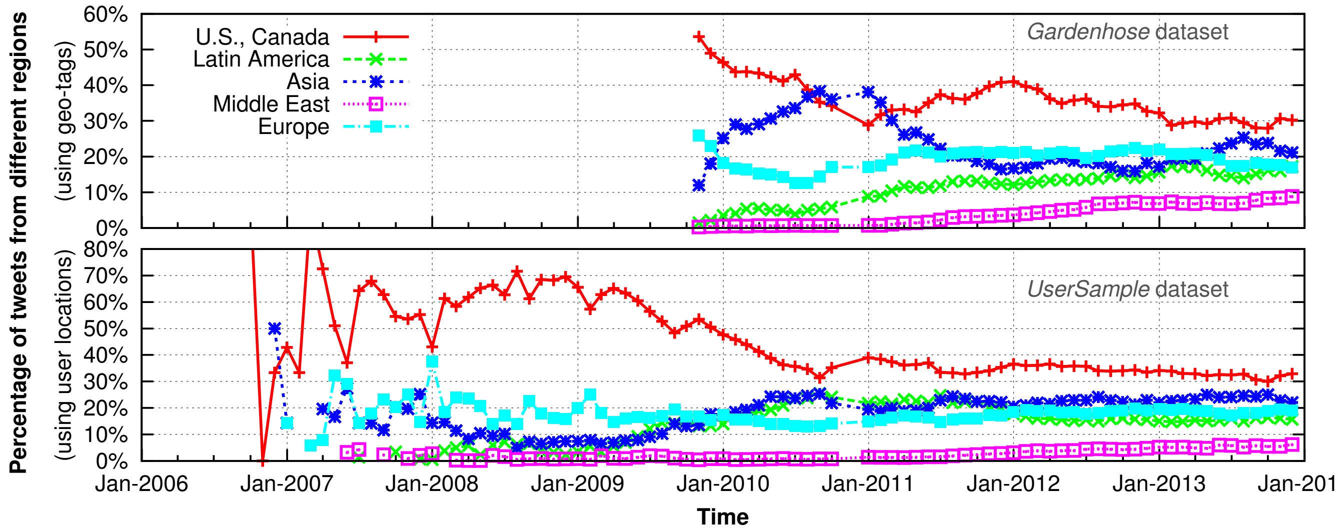

Figure 5: The percentage of tweets created in different geographical regions over time. Shown are locations inferred from self-reported user locations (*UserSample* dataset, bottom) and geo-tags (*Gardenhose* dataset, top). Geo-tags were introduced in November 2009, so we only show data since then. Africa and Oceania are both almost always less that 1%, and are not shown for clarity. A number of interesting trends can be observed, quantifying the spread of Twitter across the globe.

slowing down.

Next, we briefly examine the aggregate level of user activity as Twitter has evolved. To do so, we use observed tweets for each user in the *Gardenhose* dataset and use the first and last `statuses_count` field in the user profile (we note that using the *Gardenhose* dataset introduces a bias towards more active users). The median value of the number of tweets per user per month is presented in Figure 4. We can see a general rise in activity from later 2009 to 2012, with the rate stabilizing and then decreasing.

Users leaving Twitter While Twitter has seen massive growth, a non-trivial percentage of users leave Twitter, either by deactivating their account or by being suspended by Twitter. Additionally, many users mark their account as *protected*, meaning only their approved followers can view their tweets. Finally, many users simply leave Twitter and become *inactive*, meaning they have not tweeted for over a year. In Figure 3, we use the *UserSample* dataset to plot the percentage of the entire Twitter user base whose accounts are deactivated, suspended, protected, or inactive at different points in time (e.g., in January 2013, 2% of the user population at that time were deactivated, 4% had been suspended, 5% had marked their accounts as protected, and 20% had not tweeted for over a year).[4]

We first observe the massive percentage of inactive accounts, representing up to 32.5% of all accounts by the end of 2013. The increasing nature of this trend suggests that the leveling off of active users per month (Figure 2) may soon cause the majority of accounts to be inactive. We also observe that up to 15% of users who joined in Twitter before December 2007 protected their accounts, while the percentage goes down to 4.8% by 2013; this implies that most users who have joined Twitter recently have kept their accounts public. We further observe a dramatic increase in the percentage of suspended users, with over 6% of the entire Twitter user population suspended by late 2013; this is inline with studies on the rise of malicious activity (Thomas et al. 2013; Yeung 2013). Finally, a relatively stable 2% of users who have deactivated their accounts, indicating that relatively few users are leaving Twitter by explicitly deleting their accounts.[5]

User location Now, we examine the geographical distribution of the users over time. To do so, we rely on two pieces of information: (a) the self-reported, unformatte t the location field, we query Bing Maps with each unique location string, and only consider results that Bing returns with "high confidence". To interpret the geo-tags, we use country GIS shape files to translate latitude/longitudes into countries and administrative districts. Since geo-tags were introduced in November 2009, we only report geo-tags for months afterwards. We find 42.4% of users provide a location string interpretable by Bing, and 1.23% of tweets have included geo-tags.

We present the results of this analysis in Figure 5, showing the percentage of users located to different regions of the world using both self-reported locations (bottom) and geo-tags (top). We observe a number of interesting trends that

[4]Since Twitter does not provide the date a user's account becomes unavailable, we define the date of being suspended, deactivated, protected as the last date on which we observed a tweet from the user.

[5]We may underestimate the percentage of users who deactivated their accounts, as we can only infer that deactivated `user_ids` ever existed if we observed a tweet from them.

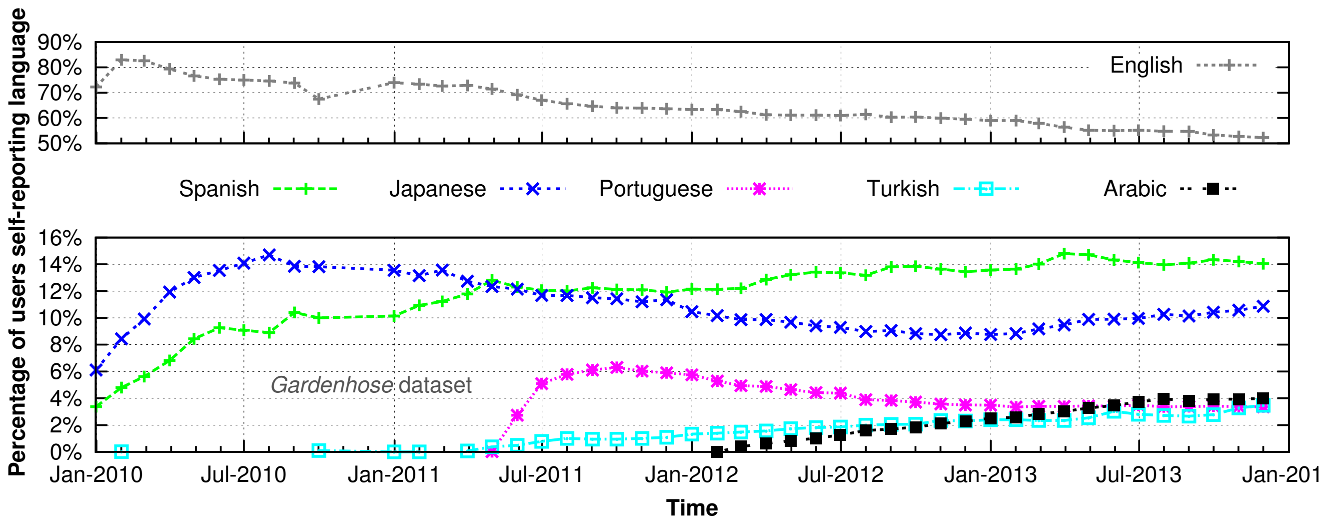

Figure 6: The percentage of users self-reporting the six most popular languages over time. We plot English separately in the top graph in order to increase the readability; note that the scale is different between the two graphs. English shows a mostly linear decrease from 83% in January 2010 to 52% in December 2013.

quantify the spread of Twitter across the globe: First, we observe a steep decline of the percentage of the tweets from the U.S. and Canada from a high of above 80% to 32%; most of this decline comes in 2009 and 2010. At the same time, we observe a substantial increase in the percentage of tweets from the Middle East (starting in early 2011, corresponding to the Arab Spring) and Latin America. However, Europe is relatively stable over the course of Twitter's evolution, generally representing around 20% of the tweets. Finally, comparing the two graphs, we can observe a difference in the spread of geo-tags, with Asia users being over-represented relative to the entire user population in 2009 and 2010. This is likely due to the popularity of smartphones equipped with GPS in Asia, which have since become popular globally.

Languages We now examine the languages Twitter users report over time. To do so, we rely on the self-reported `lang` field that Twitter allows users to specify in their profile; this field first appeared on January 12, 2010, so we report data since then in Figure 6 (English in the top graph, other languages in the bottom graph). We observe a steady (and continuing) decrease of users reporting English, from a high of 83% to 52% in December 2013. Spanish and Japanese show a steady representation of approximately 10%, while the decrease of English is correlated with an increase in a large number of other languages, including Turkish, Portuguese, and Arabic (correlating well with the results in the previous section showing the distribution of user locations). Overall, these results underscore the fact that Twitter's user population is continuing to become more diverse and global.

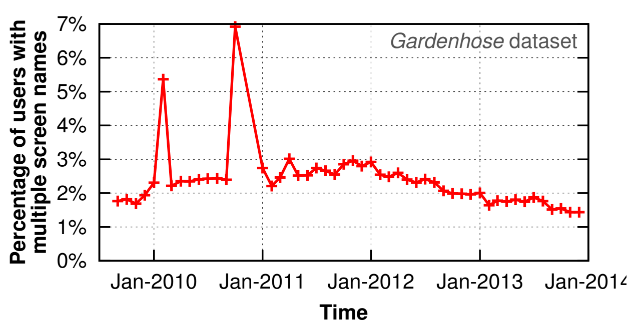

Figure 7: The percentage of users who have used more than one screen names in each month. The "spikes" correspond to times where Twitter released old, inactive screen names to be reclaimed.

Screen name changes A little-known feature of Twitter is that users can easily change their screen name (e.g., changing @Barack to @BarackObama), meaning tweets from the same user may show up under different screen names.[6] Using our *Gardenhose* dataset, we can observe these changes happening by looking for tweets from the same `user_id` with different screen names. Figure 7 plots the percentage of the user population that we observe to have used multiple screen names each month; the "spikes" in February and October 2010 correspond to time periods where Twitter opened up old, inactive screen names to be reclaimed by active users (Bryant 2010). Previous results have suggested that these users are more likely to be spammers (Chowdhury 2010), and our results correlate with prior studies showing an increase in the level of Twitter spam in 2010 and 2011 (Thomas et al. 2011; Acohido 2010). Overall, our results show that a non-trivial percentage (up to 3%) of users

[6]If a user changes their screen name, their `user_id` remains the same; this allows us to track screen name changes. There is no limit to the number of times a user can change their screen name.

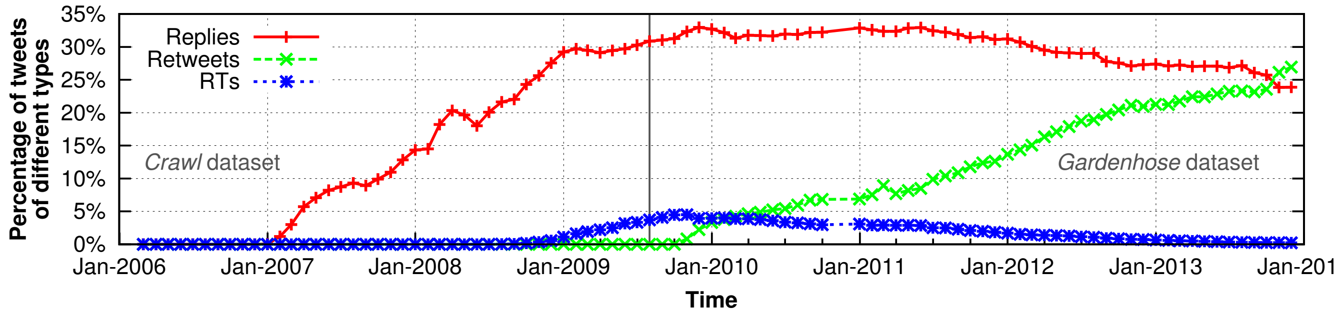

Figure 8: The percentage of tweets of different types over time. Both *RTs* and *Retweets* represent retweets, with the former capturing manually created retweets by users. Native retweets were supported by Twitter starting in November 2009.

change their screen names each month; this suggests that researchers should internally refer to users with Twitter's (unchangeable) `user_ids` to ensure that users are accurately tracked over time.

Social characteristics We now turn to examine the social characteristics of Twitter users. Recall that Twitter users can *follow* each other, and following a user means the followed user's tweets will show up when you log in. Following a user only requires permission if the followed user's profile is protected.

Using the *Gardenhose* dataset, we calculate the median number of followers (i.e., those following a user) and friends (i.e., those who a user follows) for all observed users over time, and present the results in Figure 9 (top). Similar to Figure 4, these results are biased towards more active users due to the use of the *Gardenhose* dataset. We observe a dramatic increase in the median followers/friends count of almost 400% from 2009 to 2013. This trend underscores Twitter's importance as a information dissemination platform;

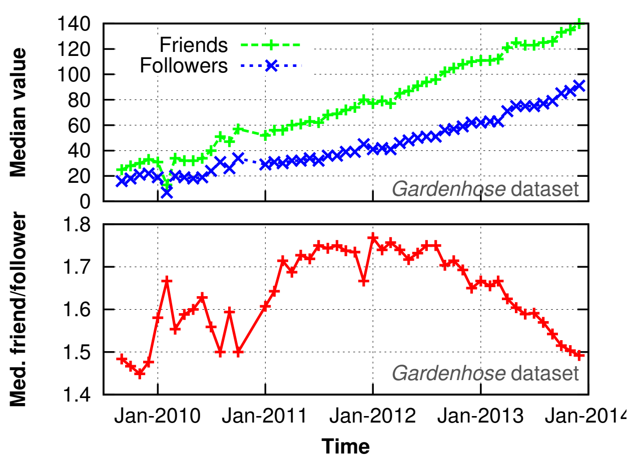

Figure 9: The median number of friends and followers across all users (top), and the median ratio of friends to followers (bottom) as derived from the *Gardenhose* dataset. We observe a dramatic densification of the Twitter social graph. Similar to Figure 4, using the *Gardenhose* dataset causes a bias towards more active users.

today, many celebrities, companies, and organizations use Twitter as one of the primary mechanisms to communicate with others (Christoforos 2011).

We also examine the average ratio of friends-to-followers in Figure 9 (bottom), and find an interesting trend: the ratio increases from 1.50 to a high of 1.77 in January 2012 before returning to its previous value. We make two observations. First, the fact that the ratio is much higher than 1 indicates that the distribution of followers is much more biased than the distribution of friends (i.e., most users have many more friends that followers) indicating that Twitter is disassortative; similar observations have been made about Twitter (Cha et al. 2010) and other social networks (Mislove et al. 2007). Second, the increase corresponds well with the rise of Twitter follower spam in 2010 and 2011 (Stringhini et al. 2012); we posit that the subsequent decrease is likely due to Twitter's more active role in suspending and deleting malicious accounts (Thomas et al. 2013; 2011).

Tweeting behavior

In the above section, we examined the changing patterns in Twitter user population, now we turn to take a look at the changes in users' tweeting behavior over time.

Tweet causes We begin by examining internal-to-Twitter actions that cause tweets, focusing on two mechanisms: retweets (i.e., a user re-sharing one of his friends' tweets with his own followers) and replies (i.e., a user replying to a tweet authored by one of his friends). While the Twitter API allowed users to create replies natively starting in early 2007, Twitter did not natively support creating retweets until November 2009 (Meeder et al. 2010; Parr 2009). Instead, users who wished to retweet a tweet manually copied the tweet and added a "RT @username" at the beginning to indicate a retweet. As a result, when calculating retweets, we need to look for both native retweets (*Retweets*) as well as manual retweets (*RTs*).

Using the *Crawl* and *Gardenhose* datasets, we calculate the percentage of tweets that are replies, retweets, and RTs over time in Figure 8. We observe a number of interesting trends. First, we can see a rapid adoption of the reply mechanism, peaking at almost 35% of all tweets in 2010 and declining slightly afterwards. Second, we observe that retweets

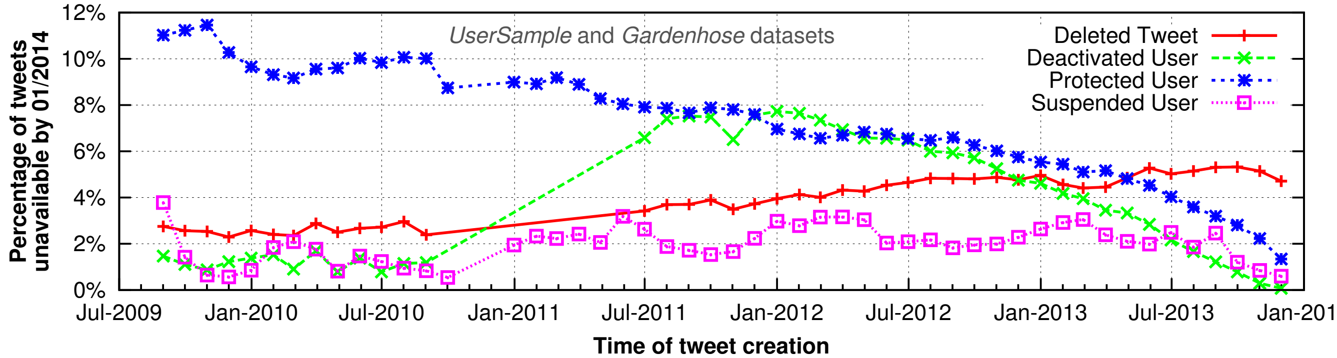

Figure 10: The percentage of tweets in the *Gardenhose* dataset that are unavailable as-of January 2014, considering only tweets issued by users in our *UserSample* dataset. Up to 10% of tweets are issued by users who later change their account to be protected, and up to 5% of tweets are explicitly deleted by users. The rise of tweets unavailable due to suspended and deleted users corresponds strongly with the increase of spam on Twitter (Thomas et al. 2011).

are initially a small percentage of all tweets, presumably due to the manual effort required to create a retweet before November 2009. However, the percentage of native retweets increases rapidly afterwards, likely due to the native retweet support that many Twitter clients provide. In fact, in late 2013, the percentage of retweets is larger than the percentage of replies. Overall, the decline in replies indicates that there is declining person-to-person communication on Twitter, suggesting significant changes in users' tweeting behavior.

Unavailability of tweets Twitter's Terms of Service (Twitter 2012) requires that any data shared about tweets is shared only in the form of a `tweet_id`; the recipient then must query Twitter to obtain the actual tweet data. Presumably, this policy is in-place so that Twitter can respect users' privacy by preventing further access to tweets that a user deletes or marks as protected. This policy significantly impacts researchers, however, as researchers wishing to reproduce prior findings may not be able to obtain the entire data sets used by others.

In order to understand the impact that this policy has, we study the percentage of tweets that become unavailable over time. In general, there are four mechanisms that could lead a public tweet to later be unavailable: (1) the tweet could be explicitly deleted by the user, (2) the user could switch their account to be "protected", thereby making their tweets only available to their approved followers, (3) the user's account could be suspended by Twitter, and (4) the user could deactivate their entire account. To study how tweets become unavailable, we need to re-query Twitter for the current status of users; since we cannot query Twitter for all 388 million users, we instead perform this analysis with the *UserSample* dataset.

In Figure 10, we present the percentage of tweets created over time that are no longer available as-of January 2014 (i.e., we look for tweets in the *Gardenhose* dataset issued by users in the *UserSample* dataset). We observe that up to 20% of these users' publicly issued tweets, in aggregate, can no longer be accessed; this significantly impacts the ability for researchers to reproduce prior results. We also observe a number of other trends. First, we observe a natural decline in the percentage of tweets unavailable due to protected and deactivated users; this is expected, as more recent tweets have had less "time" for the issuing user to become protected or deactivated. Second, we observe that the percentage of tweets that are unavailable due to suspended users is fairly constant around 2%; this suggests that Twitter quickly suspends malicious users, so time passed is not significant factor in the likelihood of a tweet being unavailable due to suspension. Third, as has been show in previous work (Almuhimedi et al. 2013), most deleted tweets are deleted quickly after being issued. The increase in the percentage of deleted tweets therefore suggests a change in user behavior over time, with users becoming more likely to delete tweets.

Tweet contents Next, we take a closer look at the contents of tweets. Examining the contents of over 37 billion tweets is quite challenging, so we focus on three types of internal tweet *entities* that Twitter natively supports: tweets with URLs (i.e., a user sharing a link), tweets with hashtags such as `#fail` (i.e., a user stating the topic of the tweet), and the tweets with mentions[7] such as @BarackObama (i.e., a user mentioning another user in the tweet). Similar to the RT syntax for retweets, the syntax for specifying both hashtags and mentions was created by the Twitter users themselves, and only later natively supported by Twitter (Cooper 2013; Stone 2009).

Figure 11 (bottom) presents the percentage of tweets that have at least one of the various entity types over time. We can observe that both mentions and URLs were popular by 2009, but that hashtags only began appearing in more than 10% of tweets in 2010. Surprisingly, since 2009, the percentage of tweets with mentions has increased substantially, while the percentage of tweets with URLs has actually de-

[7]Note that replies are by definition a type of mention, as the replying user includes the username they are replying to in their reply.

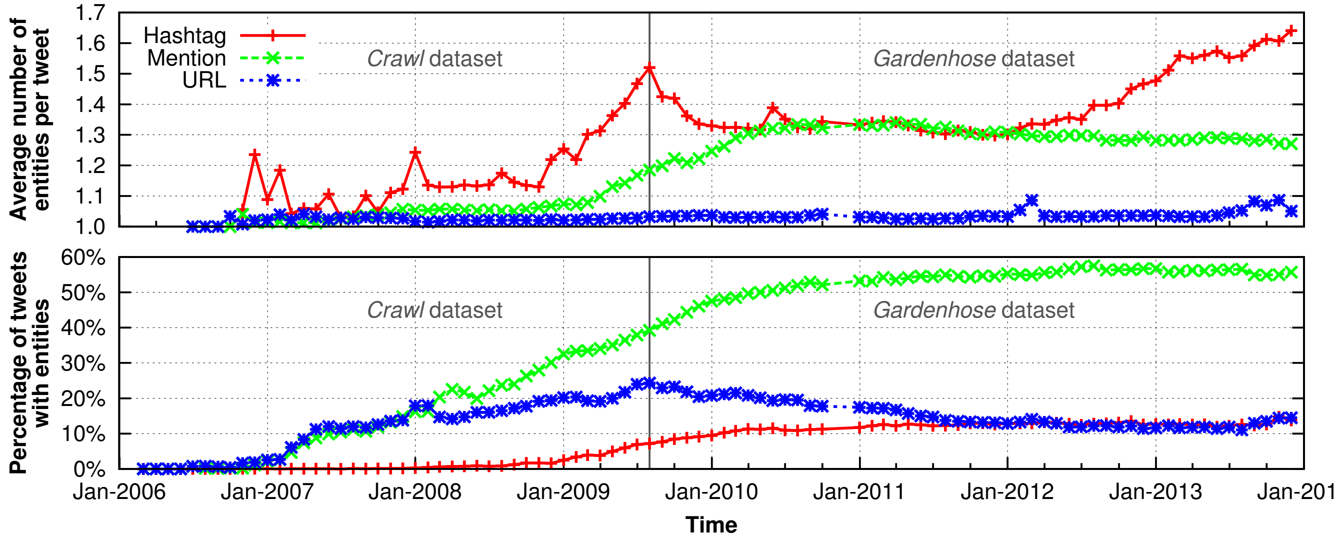

Figure 11: The percentage of tweets with different types of entities (bottom), and average number of entities for such tweets (top) over time. We observe increasing adoption of mentions, and an increasing likelihood of many hashtags per tweet.

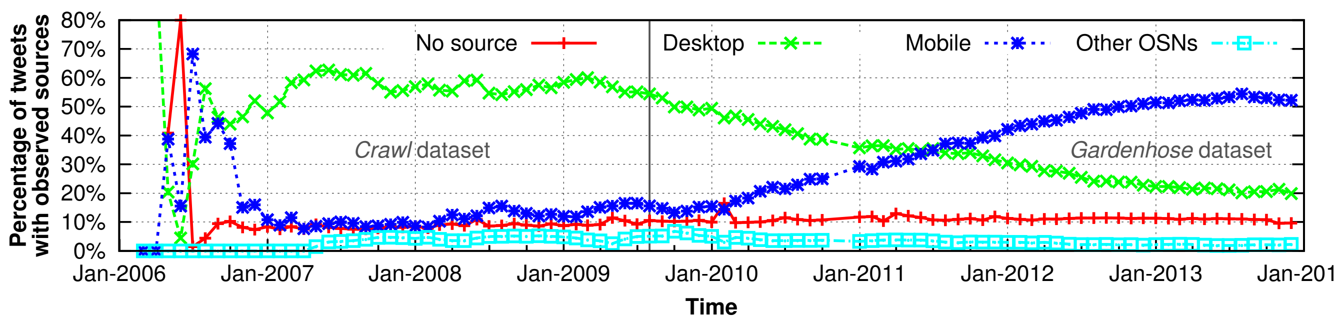

Figure 12: The percentage of tweets created with different sources (i.e., different clients) over time.

creased to stabilize at ∼12%. Overall, these results suggest that Twitter has become more "conversational", with users mentioning other users in over 50% of tweets today.

We note that users can choose to include more than one of a given entity type in a tweet (e.g., a single tweet can include multiple hashtags). To understand this behavior, Figure 11 (top) plots the average number of entities in tweets that have at least one such entity (i.e., for the hashtags line, we only consider tweets with at least one hashtag). We observe that URLs and mentions have largely stabilized around 1.0 and 1.3, respectively, but that the average number of hashtags shows a continuing increase beyond 1.6. This trend is likely explained by an increasing level of hashtag spam that has been observed (Vaas 2013; Ostrow 2009), where malicious users issue tweets with many hashtags in an attempt to make the hashtags appear popular.

Twitter clients Twitter was originally designed to be used on mobile devices by sending SMSes (hence the 140 character limit). From 2006 to the present, we have witnessed an explosion of popularity of smartphones and other mobile devices like iPads; using Twitter applications is now a popular activity on these devices. Additionally, over this time period, Twitter has become closely intertwined with other social networks such as Facebook, with many users automatically cross-posting their updates between multiple sites. As a final experiment, we explore how tweets are created by taking advantage of the source field that Twitter attaches to each tweet.

The source field is different for each different Twitter client, so we begin by manually classifying all 54 unique sources that represented at least 1% of tweets in any month. We classify the sources in categories: Desktop, Mobile, Other OSNs. Certain sources exist in multiple categories (e.g., Echofon has both mobile applications and desktop applications), so we do not include these.

We present the breakdown of tweets from different sources in Figure 12, and make a number of observations. Overall, there is a consistently decreasing trend for desktop clients (including web, API, WebClient and other applications for the desktop computers), and a corresponding increasing trend for mobile clients (including iPhone, Android, BlackBerry, iPad, and other applications for the mobile devices). We also observe that mobile devices are briefly a popular source in 2006 (using SMS), but quickly drop before rising again. These trends quantify the shift towards mobile devices, with mobile devices representing the majority of tweets starting in 2013. Surprisingly, we observe that tweets automatically created by other OSNs (including TwitterFeed, Facebook, and Tumblr) consistently represent approximately 3% of the overall tweet volume.

Related Work

In this section, we briefly detail previous studies of Twitter users and their behavior.

Sample Coverage Most studies to understand user activities on Twitter utilize the public Twitter API to collect the information of users and tweets. To measure the limitations and representation of sample data, recent work (Morstatter et al. 2013) examined how Twitter selected tweets to return, and found that many API queries are not representative. However, we use the Twitter gardenhose in our study, which Twitter explicitly states is a random selection of public tweets.

By comparing US Twitter users to the US population using census data, previous work (Mislove et al. 2011) found that US users are a highly non-uniform sample of the population in terms of race/ethnicity, gender, and geographical distribution. Our work here is complementary, as we are examining the evolution of the Twitter userbase.

Tweeting Behavior A number of researchers have examined different aspects of the user tweeting behavior. To understand why users retweet, some papers analyze the patterns and causes of retweeting, suggesting that the primary motivations for retweeting are new information for the user (Macskassy and Michelson 2011) and statements of support (Recuero, Araujo, and Zago 2011). Our results do help clarify some prior work: for example, Macskassy et al. (Macskassy and Michelson 2011) report that 32% of tweets are retweets, contradicting our measurement of 10% at the same time. The mismatch is likely caused by the authors' snowball sampling method.

Several studies have focused on the deletion of tweets. Tweets are deleted for numerous reasons; for example, they can be spam, or need to be expressed in a different way, or can cause the user regret later on (Petrovic, Osborne, and Lavrenko 2013). The most comprehensive study found (Almuhimedi et al. 2013) that the majority of tweets get deleted within an hour, and the fastest tweets to be deleted generally have typos or need rephrasing. Both studies find that around 2–3% of tweets were deleted in their 2012 dataset, which is consistent with our results (2.35%) for the same time period.

Recently, researchers have worked to gain a better understanding of how users socialize with others in the network: the conversation generated by mentioning another user (Macskassy 2012), and the top trending topics generated using hashtag (Huang, Thornton, and Efthimiadis 2010; Yang et al. 2012).

Today, people can tweet by texting, use a mobile application, use an agent application to tweet (either non mobile or mobile), among other ways. By tracking the `source` of tweets, prior work (Perreault and Ruths 2011) has shown that mobile Twitter users are more likely to be active than non-mobile users, and the tweets made on a mobile device tend to be more conversational and personal. In our paper, we find an interesting similar trending that texting (one of the few ways non-smartphone users can access Twitter) has decreased in popularity steadily since 2006; whereas smartphone devices and applications have become very popular.

User Demographics Geo-locating users has become a prominent area in the study of Twitter data. With regard to the `location` field in the user profile, prior work (Hecht et al. 2011) found that 34% of Twitter users had entered fake locations in their profile that provided no geographical information or location and 11.5% entered geotags; the rest entered valid geographical data. Our results are consistent with this study, but more comprehensive (as we shown the evolution of these trends, instead of a single snapshot).

In terms of users' self-reported `lang`, our findings supports the previous findings by Krishnamurthy (Krishnamurthy, Gill, and Arlitt 2008) about the top 10 languages on Twitter in 2008. However, we also show that this situation has changed significantly in the intervening time, with English today covering barely half of the user population.

Prediction of User Characteristics While it is largely orthogonal to our work, there has been significant work in prediction of user profile fields, including user's location, gender, race/ethnicity, and age (Mislove et al. 2011; Pennacchiotti and Popescu 2011a; 2011b; Bergsma et al. 2013; Nguyen et al. 2013). For example, recent work (Bergsma et al. 2013) found that the user-provided last name, first name, and location can provide information on the potential country of origin, the language, gender, ethnicity, and race with between 83% and 90% accuracy. Others (Nguyen et al. 2013) have shown that the specific language, sentiment, linguistic style of the tweets, as well as the tweeting behavior (retweets, hashtags) can give some idea about the age or interests of the user.

Researchers have examined how to infer the user's location from the tweet contents. For example, some studies (Hecht et al. 2011; Chandra, Khan, and Muhaya 2011; Cheng, Caverlee, and Lee 2010) showed how to predict the user's location based on the tweet contents, by looking for location-specific words in their tweets. As a result, as the amount of tweets increases, the accuracy of correctly predicting the user's location is likely to increase as well.

Concluding discussion

Twitter has seen significant interest from the research community over the past few years. However, most studies have used Twitter data as a means to an end, such as to predict aspects of the real world or to better understand information flow through society. There has been relatively little work that has studied the evolution of Twitter itself. Given that Twitter has changed significantly, it becomes unclear how to interpret prior results and whether the assumptions made in the past are still valid.

Using a dataset of over 37 billion tweets from seven years, we presented a close examination of the evolution of the Twitter user population and their behavior. We are able to strongly quantify a number of trends, including the spread of Twitter across the globe, the shift from a primarily-desktop to a primarily-mobile system, the rise of malicious behavior, and the changes in users' tweeting behavior. We hope that our findings will aid researchers in understanding the Twitter platform and interpreting prior results. As Twitter continues to develop, we plan to repeat our analysis to study the future evolution of the Twitter user population and behavior.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported in part by NSF grants CNS-1054233 and CNS-1319019, ARO grant W911NF-12-1-0556, and an Amazon Web Services in Education Grant.

References

Acohido, B. 2010. Facebook, Twitter see more spam attacks. `http://usat.ly/1gKhYiY`.

Almuhimedi, H.; Wilson, S.; Liu, B.; Sadeh, N.; and Acquisti, A. 2013. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *CSCW*.

Bergsma, S.; Dredze, M.; Durme, B. V.; Wilson, T.; and Yarowsky, D. 2013. Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter. In *NAACL*.

Bryant, M. 2010. Twitter goldrush on the way. Old usernames to be released. `http://tnw.co/1ozpteR`.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*.

Chandra, S.; Khan, L.; and Muhaya, F. B. 2011. Estimating Twitter User Location Using Social Interactions-a Content Based Approach. In *SocialCom*.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In *CIKM*.

Chowdhury, A. 2010. State of Twitter Spam. `https://blog.twitter.com/2010/state-twitter-spam`.

Christoforos, A. 2011. All these celebrities on Twitter. `http://exm.nr/1epC2X5`.

Cooper, B. B. 2013. The surprising history of Twitter's hashtag and 4 ways to get the most out of them. `http://bit.ly/1euTB8z`.

Gilbert, E., and Karahalios, K. 2010. Widespread Worry and the Stock Market. In *ICWSM*.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *CHI*.

Huang, J.; Thornton, K. M.; and Efthimiadis, E. N. 2010. Conversational Tagging in Twitter. In *HT*.

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about Twitter. In *WOSN*.

Macskassy, S. A., and Michelson, M. 2011. Why do People Retweet? Anti-Homophily Wins the Day! In *ICWSM*.

Macskassy, S. A. 2012. On the Study of Social Interactions in Twitter. In *ICWSM*.

Mao, H.; Shuai, X.; and Kapadia, A. 2011. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *WPES*.

Meeder, B.; Tam, J.; Kelley, P. G.; and Cranor, L. F. 2010. RT@IWantPrivacy: Widespread violation of privacy settings in the Twitter social network. In *W2SP*.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and Analysis of Online Social Networks. In *IMC*.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the demographics of Twitter users (Poster Paper).

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. How Old Do You Think I Am? A Study of Language and Age in Twitter. In *ICWSM*.

Ostrow, A. 2009. Twitter Spam Invades Trending Topics. `http://on.mash.to/1dfFANu`.

Parr, B. 2009. PROJECT RETWEET: Twitter's Plan to Officially Support Retweeting. `http://on.mash.to/1fJK9tQ`.

Pennacchiotti, M., and Popescu, A.-M. 2011a. A Machine Learning Approach to Twitter User Classification. In *ICWSM*.

Pennacchiotti, M., and Popescu, A.-M. 2011b. Democrats, republicans and starbucks afficionados: user classification in twitter. In *KDD*.

Perreault, M., and Ruths, D. 2011. The Effect of Mobile Platforms on Twitter Content Generation. In *ICWSM*.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2013. I Wish I Didn't Say That! Analyzing and Predicting Deleted Messages in Twitter. *CoRR*.

Post, H. 2014. Twitter Statistics. `http://bit.ly/18KIwd2`.

Recuero, R.; Araujo, R.; and Zago, G. 2011. How Does Social Capital Affect Retweets? In *ICWSM*.

SEC. 2013. Twitter, Inc. Common Stock. `http://1.usa.gov/1gVlt2o`.

Stone, B. 2009. Replies Are Now Mentions. `https://blog.twitter.com/2009/replies-are-now-mentions`.

Stringhini, G.; Egele, M.; Kruegel, C.; and Vigna, G. 2012. Poultry Markets: On the Underground Economy of Twitter Followers. In *WOSN*.

Thomas, K.; Grier, C.; Song, D.; and Paxson, V. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *IMC*.

Thomas, K.; McCoy, D.; Grier, C.; Kolcz, A.; and Paxson, V. 2013. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. In *USENIX Security*.

2012. Twitter Terms of Service. `https://twitter.com/tos/`.

Vaas, L. 2013. Spammers take over Twitter Trends with sexy hashtags. `http://bit.ly/OArYAX`.

Weil, K. 2010. Measuring Tweets. `https://blog.twitter.com/2010/measuring-tweets`.

Yang, L.; Sun, T.; Zhang, M.; and Mei, Q. 2012. We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption? In *WWW*.

Yeung, K. 2013. Spammers are back on Twitter, and this time they're coming after you through lists. `http://tnw.co/1epCTqJ`.