# Tweets Are Forever: A Large-Scale Quantitative Analysis of Deleted Tweets

**Hazim Almuhimedi[a], Shomir Wilson[a], Bin Liu[a], Norman Sadeh[a], Alessandro Acquisti[b]**
[a]School of Computer Science, [b]Heinz College
Carnegie Mellon University
{hazim,shomir,bliu1,sadeh}@cs.cmu.edu, acquisti@andrew.cmu.edu

## ABSTRACT

This paper describes an empirical study of 1.6M deleted tweets collected over a continuous one-week period from a set of 292K Twitter users. We examine several aggregate properties of deleted tweets, including their connections to other tweets (e.g., whether they are replies or retweets), the clients used to produce them, temporal aspects of deletion, and the presence of geotagging information. Some significant differences were discovered between the two collections, namely in the clients used to post them, their conversational aspects, the sentiment vocabulary present in them, and the days of the week they were posted. However, in other dimensions for which analysis was possible, no substantial differences were found. Finally, we discuss some ramifications of this work for understanding Twitter usage and management of one's privacy.

### Author Keywords

Privacy; Social Networks; Deletion

### ACM Classification Keywords

H.3.1. Information Systems: Information Storage and Retrieval: Content Analysis and Indexing

## INTRODUCTION

Online social networks such as Twitter, Facebook, Google+, and LinkedIn have established themselves as channels for a variety of communications between individuals. Archival formats in social networks, known by terms such as *news feeds* or *timelines*, often allow the users to view the histories of their peers' posts. Since one's communication history is on display for others to browse or scrutinize, it is natural that online social networks often (or almost always) include some mechanism for users to delete posts that they previously made. Reasons to do this are numerous: posts may include typos, factual errors, stale information, reconsidered ideas, or regretted statements. Sometimes deletions are superficial in nature, but in other cases they may have serious ramifications, as recognized by the European Commission's draft of a "right to be forgotten" [1].

When a post is deleted from an online social network, users generally assume that the post will no longer be available for anyone to see. However, this is not necessarily true, as evidence may persist of the post and its content in less visible ways. Twitter, through its API service, provides a particularly rich and accessible stream of data on deleted posts. By following the posts (*tweets*) of a user and other messages from the API, one can reconstruct which tweets the user decides to delete without losing any data associated with them. By tracking a large number of users whose posts are public, it is thus possible to observe large-scale patterns in deletion behavior. These patterns can inform the design of online social networks to help users better manage their content.

The fact that Twitter users sometimes delete tweets seems intuitive—if not obvious—but the ramifications of deletion have received little attention. Specifically, we wish to address the following questions:

- *What differences exist between deleted and undeleted tweets?* For instance, are certain properties of tweets (e.g., location tags, reply frequencies) correlated with likelihood of deletion? Might there be patterns in sentiments or topics among deleted tweets?

- *How often are tweets deleted for superficial reasons?* These include typos, simple rephrasings, spam generators, and mass deletions. Some superficial deletions may represent a waste of effort that could be prevented with enhancements to Twitter clients, while others may involve little or no action on the part of a human user.

- *What are some of the additional reasons that people delete tweets?* Answering the above two questions may provide some insight (albeit indirect) into user intentions. A comprehensive account of reasons for deletion might be impossible, given the challenges of reconstructing users' motivations and specific circumstances. However, some hypotheses can be formulated by examining the properties of deleted tweets *en masse*.

This paper describes the first large-scale empirical study of deleted tweets gathered from the Twitter online social network. A corpus of approximately 1.6M deleted tweets was


*CSCW '13,* February 23–27, 2013, San Antonio, Texas, USA.

gathered over a continuous one-week period from a set of approximately 292K users who made their posts available to the general public. The same users' persistent undeleted tweets also were gathered, providing a comparative set. Several features of the two sets of tweets were contrasted, including the source clients, tweet content, quantity of replies, and the presence of geotagging information. The researchers found that many of the properties of deleted tweets in aggregate differed from undeleted tweets, while others were indistinguishable. Finally, some implications of these findings are discussed, along with some directions for future work.

## TWITTER

The reader is likely to be familiar with Twitter already, but for completeness, below is an overview of the service and its terminology.

Twitter is an online social network that allows users to post 140-character long messages, known as *tweets*. A user can *follow* other users (sometimes informally referred to as their *friends*, as in many similar networks) to easily receive their tweets in an aggregated news feed. By default, all tweets are public and accessible by anyone who browses Twitter's website. However, a user can *protect* their account and only allow a selected group of approved followers to view their tweets. For this study, the researchers only gathered data from public Twitter accounts.

A Twitter user can engage in a conversation by replying to a tweet posted by another user or by *mentioning* a user; both methods use the convention of including "@username" in the reply tweet or mentioning tweet. Additionally, a Twitter user can *retweet* another user's tweet if it is public, thus passing it along to her followers. To assign a topic to a tweet, Twitter users follow the convention of using *hashtags* (e.g., "#topic"). Tweets frequently contain URLs to websites or images, and they are sometimes tagged with location metadata.

### Tweet Deletion

Twitter users are able to delete tweets that they have posted. Upon deletion, the tweet disappears from the user's timeline and from their follower's timelines as well. A deleted tweet effectively disappears from the results of searching Twitter, although a short delay sometimes occurs between deletion and disappearance. A *status deletion notice*[1] is distributed via the Twitter streaming API to relevant users' clients so that they, in turn, remove deleted tweets from their records. A user also may delete a retweet through an undo mechanism, although this does not delete the original retweeted post, since it belongs to a different user. On the other hand, if the author of a tweet that has been retweeted deletes it, then all retweets are deleted as well.

Twitter does not provide a bulk-deletion of user's tweets. It provides, however, a one-click bulk-deletion of all location data that were attached to user's tweets, without deleting the tweets. By clicking on the "Delete all location information" button on user's account settings page, all locations attached to all previous tweets are deleted. A "location deletion notice", which includes a "user ID" and an "up-to-status ID", will be sent to third party clients so they, in turn, can delete the location information of tweets from their back-end data stores.

[1]https://dev.twitter.com/docs/streaming-apis/messages

## RELATED WORK

A large volume of research has been conducted on Twitter, including social network topology and properties [13, 11] as well as user classifications, behaviors, and topics of tweets [27, 26, 11, 10, 12]. The contents of tweets have been a particular focus of recent research. Such efforts have involved manual coding of small samples [17], utilizing the crowd [2] and automatic topic modeling for large samples [19].

Tweet deletion has received little direct attention in previous studies, although we note one prior effort focused on this topic. Hovey [9] created a tool to recover and display deleted tweets, with statistics on their lifespans and most frequently occurring keywords. Although he explored applying natural language processing methods to deleted tweets, no results were reported. Our work differs from Hovey's in many dimensions, including the volume of the collected data, the reported comparisons between deleted and undeleted tweets, temporal and spatial analysis of deleted tweets, and the keyword-based analysis of tweet content.

Prior studies have demonstrated that users of online social networks delete their posts to manage social consequences and maintain their privacy. Wang et al. showed that Facebook users most commonly handle regrettable posts by deleting them [25]. Boyd showed that deleting posts is a "structural" strategy that teens use to minimize the social consequences of their posts, rather than taking advantage of an online social network's privacy controls [5]. Other methods of image management by teenagers include deletion of other users' posts that might cause embarrassment (e.g., making fun of one's interests) [5] and deleting posts that present a negative image of oneself to other readers [14]. Tufekci determined that 81.3% of Facebook users delete information from their profiles due to privacy or visibility concerns, and women are more likely than men to do so [24]. Some Facebook users who are concerned about privacy delete all their posts periodically [20].

Although most social networks offer deletion functionality, data is not necessarily removed immediately or erased completely. Bonneau and Cheng [4, 7, 8] separately demonstrated that deleted photos on Facebook were still accessible after the user requested their removal, almost a month and a year later, respectively. More recently, Facebook solved this problem, and deleted photos are erased within a month. Twitter deletes photos and posts instantaneously [7], but it also allows third party applications to access and archive users' tweets. No guarantee can be made that third parties will comply with users' deletions, and a few services have attempted to archive deleted tweets for public viewing. Among them were *Undetweetable* and *Tweleted*, although neither of these remains operational due to actions taken by Twitter. *Politwoops* is a recent service dedicated to storing and recovering tweets deleted by politicians. It started in the Netherlands and now

| | |
|---|---|
| Total number of users | 292,293 |
| Total number of all posted tweets | 67,295,171 |
| Total number of undeleted tweets | 65,677,499 |
| Total number of deleted tweets | 1,617,672 |
| Number of users who tweeted at least 1 tweet | 222,185 |
| Number of users who deleted at least 1 tweet | 144,816 |

**Table 1: Totalized statistics on the tweets in our dataset.**

| Statistic | Mean | Median | Std. Dev. |
|---|---|---|---|
| Followers | 1,281 | 305 | 30,114 |
| Following | 541 | 289 | 2,235 |
| Account Age (mo.) | 21 | 20 | 11 |

**Table 2: Statistics on the users in our dataset.**

has 13 different international versions including one in the U.S.

In summary, although prior research linked deletion of posts in social networks with privacy and social concerns, the phenomenon has not been examined using a volume of data comparable to the present study, nor with attention to the metadata associated with deleted tweets. We believe that this large-scale, multifaceted examination is a necessary step in exploring the ramifications of deleted user content.

## DATA COLLECTION

Using Twitter's streaming API, we first collected a random set of users who matched the following criteria: (1) their account was at least a month old; (2) the user had posted at least 10 tweets; (3) the user had at least 10 followers and 10 following; and (4) the first tweet we received through the API was in English (as determined by the Google Translate and Microsoft Translator APIs) and not a retweet. This resulted in a set of approximately 292K users, whom we followed via the streaming API for one continuous week (2012-03-14 through 2012-03-20). This allowed us to receive all public tweets posted by those users as well as retweets posted by them, replies to their tweets by their followers, and retweets of their tweets. In addition, we collected metadata for each tweet such as hashtags, URLs, user mentions, and location information. When a user deleted a tweet, a deletion notice was sent via the API containing identifiers for the user and the specific tweet. Similar deletion notices were sent for each tweet when a user removed the location information from all of their tweets. These deletion notices allowed us to divide the tweets in our dataset into those that were deleted and those that were not.

**All Users**

| Statistic | Mean | Median | Std. Dev. |
|---|---|---|---|
| Deleted Tweets | 7.2 | 1 | 43 |
| Undeleted Tweets | 296 | 168 | 374 |

**Deletion-Active Users**

| Statistic | Mean | Median | Std. Dev. |
|---|---|---|---|
| Deleted Tweets | 11 | 3 | 53 |
| Undeleted Tweets | 387 | 259 | 403 |

**Table 3: Statistics on the numbers of deleted and undeleted tweets per user in our dataset. "Deletion-active" users are those who deleted at least one tweet during the data collection period.**

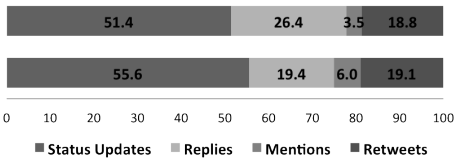


**Figure 1: Breakdown of deleted (bottom) vs. undeleted tweets (top).**

Table 1 lists some total figures for our dataset, and Table 2 shows some descriptive statistics on the included users. The mean numbers of followers and followees were skewed by very small numbers of public figures and automated accounts, respectively. In comparison to other studies of the Twitter population, our sample appears to be slightly biased toward users who have more followers and fewer followees [22, 6], perhaps because such accounts are more active and thus had posted the requisite number of tweets for our collection criterion.

## GENERAL ANALYSIS

### Breakdown of Deleted vs. Undeleted Tweets

The 67.2M collected tweets consist of approximately 65.6M (97.6%) undeleted tweets and 1.6M (2.4%) deleted tweets. Tweets were sorted into one of four mutually exclusive categories: retweets, replies, tweets that mention other users, and *status updates* (i.e., tweets that fit none of the preceding categories). Replies that mentioned other users were labeled solely as replies.

Figure 1 shows the breakdown of deleted and undeleted tweets among these categories. Compared to undeleted tweets, deleted tweets contain more status updates (55.6% vs. 51.4%), mentions (6% vs. 3.5%), and retweets (19.1% vs. 18.8%), but deleted tweets contain fewer replies (19.4% vs. 26.4%). The distributions of deleted and undeleted tweets differ significantly among the four categories of tweets ($\chi^2 = 64789.2, df = 3$, p-value $<0.05$). Pairwise per-category differences between deleted and undeleted tweets were all significant as well.[2] From a conditional probability perspective, given that a tweet is a mention or a status update, its likelihood of being deleted is slightly higher than if it is a reply or retweet.

### Deleted vs. Undeleted Tweets Per User

Not all users had tweeted or deleted a tweet during the one week period. Table 3 presents some statistics on the quantities of deleted and undeleted tweets per user. Of those 292,293 Twitter users we tracked, 76% had posted at least one tweet during the monitoring period, and 49.5% deleted at least one tweet during the same period. Out of all users who posted at least once, only 65% of them deleted at least one tweet during the same period. Figures 2 and 3 show the distributions

[2]We applied Bonferroni correction for all per-category (i.e. multiple comparison) statistical tests throughout this paper.

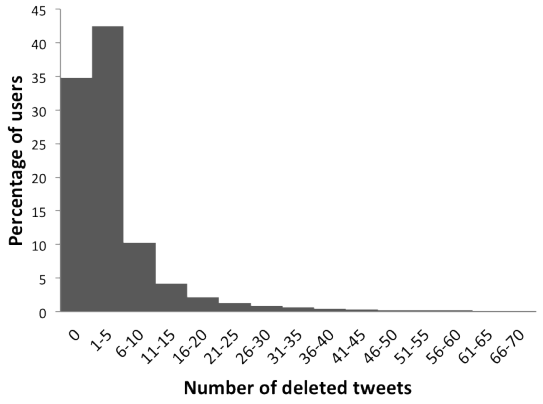


**Figure 2: The percentage of users per number of deleted tweets.**

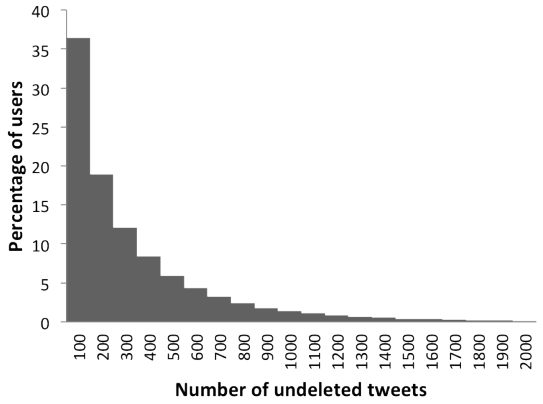


**Figure 3: The percentage of users per number of undeleted tweets.**

of users across quantities of deleted and undeleted tweets, respectively.

Included in our dataset were a few unusually active users. The two top users deleted 4,848 and 3,326 tweets, respectively. The first account was a bot account, and its tweets were posted and deleted automatically.[3] All its deleted tweets had the same pattern: mentioning a number of random users with no textual content. The second account was a regular account but for a very active user. However, by excluding the deleted retweets, the number of deleted tweets for this user dropped by 71% from 3,326 to only 971.

### Sources of Deleted vs. Undeleted Tweets

Tweets are posted from a variety of official and third-party clients. In total, deleted tweets were posted from 1,252 unique clients, while undeleted tweets were posted from 3,395 unique clients. However, 99% of deleted and undeleted tweets were posted from 110 and 132 unique clients, respectively. The top five clients for both deleted and undeleted

[3]It was not possible to methodically remove bot accounts from the dataset, due to their highly varied behaviors. However, even this extremely active account produced only 0.3% of all tweets in the dataset.

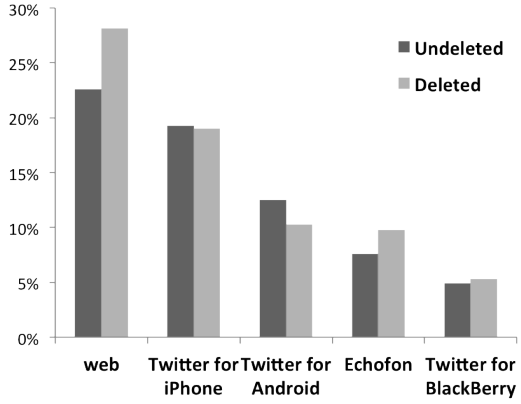


**Figure 4: Top 5 clients for deleted vs. undeleted.**

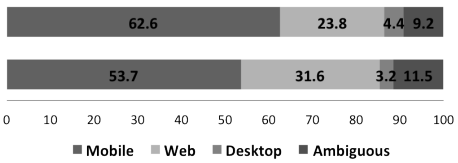


**Figure 5: Source of deleted (bottom) vs. undeleted (top) tweets.**

tweets were identical; the percentages, however, were different. Figure 4 shows the percentages of deleted and undeleted tweets from each of the top five clients. The Twitter website and the official Twitter apps dominate the sources of both deleted and undeleted tweets. The website was the source of the single largest percentage of deleted and undeleted tweets, with 28% and 23%, respectively. Only one non-official Twitter client, Echofon, was among the top five sources of deleted and undeleted tweets. Echofon has Mac, Windows, iPhone, and iPad clients as well as a Firefox extension.

We sorted the sources of 99% of deleted and undeleted tweets into web, mobile, and desktop client categories, in order to determine whether a greater proportion of deleted tweets originated from one of them. Web-originating tweets were posted from the official Twitter website, Twitter buttons on other sites, or other web clients and applications such as TwitLonger or Like My Tweet. The mobile category refers to tweets that were posted from mobile devices (including posts via mobile websites), and the desktop category refers to non-web based desktop clients. If the source of a tweet is not clear,[4] the source is labeled as ambiguous. The majority of deleted and undeleted tweets were posted from mobile or web clients. 31.6% of deleted tweets were posted from web sources, whereas 23.8% undeleted tweets were posted from web sources. 53.7% of deleted tweets were posted from mobile clients, whereas 62.6% of undeleted tweets were posted from mobile clients. Figure 5 shows the percentages of deleted and undeleted tweets for all four categories.

[4]One example is Echofon, which refers to multiple clients. All of them use an identical source tag ("Echofon") for tweets, thus making it hard to identify the tweet source.

The distributions of deleted and undeleted tweets differ significantly among the four types of Twitter clients ($\chi^2 = 74789.7, df = 3$, p-value $<0.05$). Comparing deleted to undeleted tweets in every type of Twitter client, deleted tweets were posted significantly more often from web clients ($\chi^2 = 52532.6, df = 1$, p-value $<0.05$) and significantly less often from web ($\chi^2 = 52532.6, df = 1$, p-value $<0.05$) and desktop ($\chi^2 = 5475.8, df = 1$, p-value $<0.05$) clients.

### Replies to Deleted and Undeleted Tweets

It is natural to expect that deleted tweets are less likely to receive replies, and this was found to be true. However, not all deleted tweets went unnoticed by the Twitter community. 11% of deleted tweets initiated conversations (i.e., received one or more replies) prior to their deletion. These conversations varied in message volume. The mode of the number of replies was one (covering 89% of deleted tweets that received any replies), the mean was 1.2, the median was 1, and the maximum was 524. For comparison, 27% of undeleted tweets started conversations, with a mode of one (covering 93% of undeleted tweets that received any replies), a mean of 1.1, a median of 1, and a maximum of 838. Deleted and Undeleted tweets differ significantly in the number of replies they receive ($\chi^2$, p-value $<0.05$).

Those conversations initiated by deleted tweets were rarely voluminous. The three deleted tweets that generated the most replies received 524, 263, and 176 of them, respectively. All three consisted of offers to follow users, and two of them proposed to do so in return for liking the author's tweets via third party services. Large numbers of people replied to these tweets to ask to participate. The tweet that generated the fourth most replies was a user's announcement that a university had rejected her admission application; her followers replied with sympathy and requests for more information. A full examination of the topics and contents of conversation-provoking deleted tweets was beyond the scope of this study, although it might be fruitful for future research.

To examine the connection between conversations and deletion rate, we compared the percentage of tweets with zero replies that were deleted to the percentage of tweets with one or more replies that were deleted. 2.9% of tweets with zero replies were deleted, and 1.0% of tweets with one or more replies were deleted. This difference was statistically significant ($\chi^2 = 157413.9, df = 1$, p-value $<0.0001$). This may suggest Twitter users are reluctant to delete their posts once they have received replies, or tweets that provoke replies tend to contain content that the user is less likely to delete.

## SUPERFICIAL DELETIONS

Using cues in tweet content and metadata, we identified tweets that were deleted for two particularly superficial reasons: *typos or rephrasing* and *spam*. Typos and rephrasing comprise about 17% of deleted tweets, while spam comprises 1%. Below, we describe the heuristics used to assign these labels.

| Topic | Keyword count | Keyword source |
|---|---|---|
| Alcohol and drug use | 847 | WordNet (alcohol.n.01, drink.n.02, drug_of_abuse.n.01) |
| Sexual activity | 157 | WordNet (sexual_activity.n.01) |
| Religion and politics | 273 | WordNet (god.n.01, politics.n.02, religion.n.01, religion.n.02) |
| Offensive comments | 349 | noswearing.com keyword list |

**Table 4: Selected topics and sources for keyword coverage investigation. For topics with WordNet as a source, all hyponyms were gathered for the indicated lemmas. Keyword counts include both single words and collocations (e.g., both *alcohol* and *alcoholic beverage*).**

| Topic | Tweet Coverage (%) | |
|---|---|---|
| | Deleted | Undeleted |
| Alcohol and illegal drug use | 1.69 | 1.94 |
| Sexual activity | 7.96 | 7.38 |
| Religion and politics | 0.0964 | 0.178 |
| Offensive comments | 9.97 | 10.4 |

**Table 5: Coverage by the keyword lists for selected topics over deleted and undeleted tweets.**

### Deleting for Typos and Rephrasing

Twitter users sometimes delete tweets and replace them with similar ones to correct misspellings, missing mentions, missing hashtags, or other small issues, as previously noticed by Hovey [9]. Users also may replace tweets to make changes that are semantically substantial but limited in scope (e.g., replacing a one-word obscenity with a milder term). We assigned the *typos or rephrasing* label to all tweets apparently deleted for these reasons, and detected them as follows.

If a deleted tweets is *similar* to any $K$ subsequent tweets from the same user, it is labeled a typo or replacement. We define similarity using a combination of edit distance (i.e., Levenshtein distance [15]) and cosine similarity. Edit distance is applied to detect deletions due to changes such as misspellings and missing hashtags, since it measures character-level similarity between tweets. To identify changes in word order, cosine similarity was utilized to take into consideration word-level similarity. To tweak parameters used in this procedure, we manually gathered 200 deleted tweets that contained typos or rephrasing and 200 tweets that did not. We then experimented with different distance and similarity thresholds, and determined that an edit distance of 5, a cosine similarity of 0.6, and a scanning memory of K = 3 were optimal.

### Spam

Twitter identifies twenty different classes of spamming behaviors, including aggressive following and unfollowing, using or publicizing services that sell followers, and unsolicited

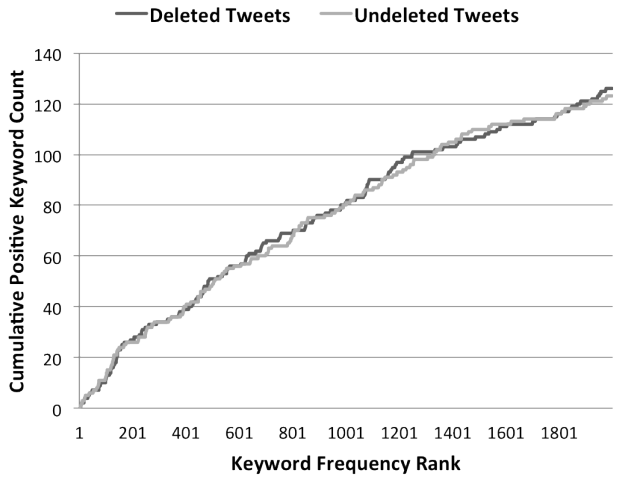


**Figure 6: Cumulative distributions of positive-sentiment keywords over frequency-ranked keywords in tweets.**

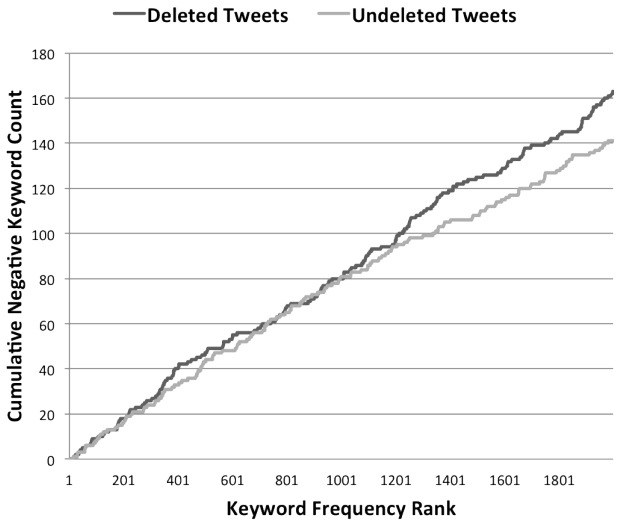


**Figure 7: Cumulative distributions of negative-sentiment keywords over frequency-ranked keywords in tweets.**

replies or mentions.[5] By focusing on those behaviors that related to the sources and contents of tweets, we were able to identify a lower bound for the percentage of spam among deleted tweets. We manually classified Twitter clients into those that produced only spam and those that did not. A client was considered a spam producer if it promoted or provided a service that claimed to give users more followers, or the client produced a large number of unsolicited replies or mentions.

Just above 1% of deleted tweets came from 14 sources that matched at least one of the above conditions. Most of these identified sources tweeted links that Twitter flagged as spam. Nearly all of the tweets from these sources (over 99%) promoted services that claimed to give users more followers, although a few were mentions and replies. Examples of spam tweets include *"GET MORE FOLLOWERS I WILL FOLLOW YOU BACK IF YOU FOLLOW ME - http://shortlink [Like it? http://shortlink]"*, *"I just gained 100+ followers from @user GO Follow her if you need more followers"*, and *"@user #hitfollowteam @user #rt @user"*. It is important to note that our approach to identifying spam among deleted tweets is a conservative one and should only be regarded as a lower bound.

### Lexical Analysis

One of our motivations for examining deleted tweets was the study of regret in online social networks, since a link between regretted content and some fraction of deleted content seems plausible. Although regrettable content is subjective (perhaps highly so) and tweets may be deleted for a variety of other reasons, it was hypothesized that some stereotypically regrettable topics might appear with greater frequency in deleted tweets than in undeleted tweets. In this section we present a preliminary evaluation of that hypothesis, as well as some related results on sentiment analysis.

[5]The Twitter Rules; `https://support.twitter.com/articles/18311`

We decided to take a lexical approach to measuring topic frequency, as more advanced methods (such as latent Dirichlet allocation [3] for topic modeling) did not produce sensible results on the available data. Four stereotypically regrettable topics were selected, as a subset of those identified from interviews with Facebook users by Wang, et al.: alcohol and illegal drug use, sexual activity, religion and politics, and offensive comments [25]. Lists of keywords and collocations were constructed for each of those topics. Table 4 shows further information on each topical keyword list. The coverage of these keyword lists (i.e., the percentage of tweets that contained at least one word from a given list) was then calculated for approximately 10,000 randomly chosen deleted and undeleted tweets, respectively. Words in tweets were stemmed prior to the coverage calculations.

Table 5 shows the coverage of each keyword list on the sets of deleted and undeleted tweets. By this measure, it appears that the two sets are roughly equal in occurrences of the selected topics. Since word sense disambiguation was not performed, it initially seemed possible that certain polysemous words (e.g., "love" in the keyword list for sexual activities) had dominated coverage and obscured actual differences between the two populations. However, manually examining the top occurring keywords for each topic did not reveal such a pattern. It is either the case that more advanced methods are necessary to determine differences between deleted and undeleted tweets for these topics, or such differences are too subtle to detect.

Finally, a different keyword-based approach was used to assess differences in sentiment between deleted and undeleted tweets. The top 2,000 most frequently occurring words in deleted and undeleted tweets (respectively) were scanned for occurrences of words in AFINN-111, an annotated list of words that frequently express positive or negative sentiment in online social network posts [18]. Figure 6 and Figure 7

show cumulative distributions of positive and negative words in the frequency-ranked lists of words in deleted and undeleted tweets. Differences were expected to be subtle, since tweets widely vary in vocabulary and topics.

It appears that positive words occur with roughly the same distribution across the frequency rankings for both sets of tweets, though a slight difference appears between ranks 401 and 601. For negative words, a greater difference emerges past rank 1,200 and persists through the end of the examined frequencies. It appears that deleted tweets share much of the same sentiment vocabulary with undeleted tweets, although there is a possibility of more negative words in the *long tail* of the frequency distribution.

## TEMPORAL ANALYSIS

Tweet deletion notices do not contain timestamps that indicate when users took actions to delete tweets. We observed that deletion notices sometimes appear in the Streaming API slightly prior to the tweets they refer to, and this raised concerns about the time difference between receiving a deletion notice and the actual deletion. To determine the accuracy of the deletion timestamp as an estimate of the actual time of deletion, we conducted an experiment in which we posted 100 tweets and deleted them immediately. We recorded times for posting, deletion, and receiving deletion notices. Then, we calculated the differences between actual deletions and receptions of the deletion notices. Mindful of possible differences in responsiveness due to changing server loads and network congestion, we conducted this experiment at several different times throughout the day. We concluded that the average time difference between deleting a tweet and receiving its deletion notice is negligible.

The results of this experiment suggest that the time of receiving a deletion notice is an accurate estimate of the time of actual deletion. For clarity, the remainder of this section implicitly assumes that theory.

### How Fast Is a Tweet Deleted?

Figure 8 shows a breakdown of tweet deletion over days with a maximum of one week, as this was the duration of our data collection.[6] However, a large fraction of tweets were deleted within a much shorter period of time. About 17% of tweets were deleted within 30 seconds, 22% within a minute, 58.6% within 30 minutes, and 65.2% within an hour. On average, tweets were deleted 8.45 hours after they were posted with a standard deviation of 20.85 hours.

Although most tweets were deleted shortly after they were posted, the duration between posting and deleting a tweet differs significantly for tweets that received the two superficial deletion labels. Tweets labeled as typos and rephrasing are the fastest to be deleted, with a mean of two hours and a standard deviation of ten hours. Spam tweets are deleted, on average, five hours after being posted with a standard deviation of 11 hours. Tweets that are not spam or typos and reposts,

[6]Some tweets may have been deleted more than seven days after they were posted, but those deletions exceeded the tracking period of the study.

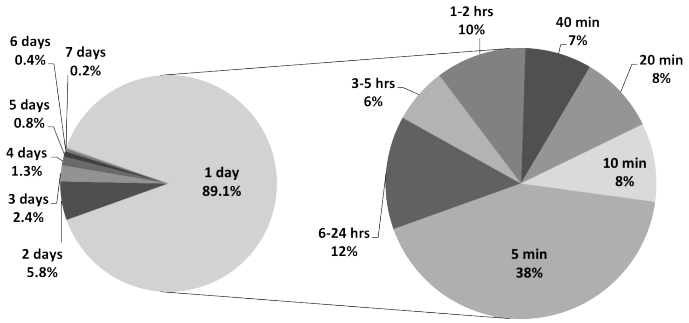


**Figure 8: How fast is a tweet deleted?**

on the other hand, are the slowest, with a mean time to deletion of 10 hours and a standard deviation of 10 hours. The differences between the lifespans of deleted tweets in these three populations are statistically significant (Kruskal-Wallis test, p-value$<$0.05).

The relationship between the posting frequency of a user and how rapidly they delete tweets is illustrated in Figure 9. Greater variation in lifespan toward the right side of the graph was the result of increasing sparsity in the data (i.e., fewer users who posted at greater frequencies). There appeared to be an inverse relationship between posting frequency and time to deletion, especially for frequencies less than 200. We observed that the same relationship appears between *deletion frequency* and time to deletion. Intuitively, this suggests that Twitter users who post (and delete) more often also respond more quickly when they sense a reason to delete a tweet.

### Analysis Over Days of the Week

We hypothesized that the rate of tweet deletion might increase on the weekends, due to social factors generally absent in the workplace. Although it was not possible to verify the motivations behind deletions, the rate was indeed greater than average on Saturdays and Sundays. On a per-day basis, Sunday had the largest percentage of deleted tweets (15.7%) and Tuesday had the smallest percentage of them (12.9%). 70% of deleted tweets were posted during weekdays, and 30% were posted during the weekend. Figure 10 shows this breakdown. Averaging over days as appropriate, the difference in the fraction of deleted tweets posted on a weekday and a weekend day was statistically significant ($\chi^2 = 924.3, df = 1$, p-value $<$0.05). Thursday was the peak day for spam, with 15.3% of the weekly total for its label, while the peak for typos and rephrasings was on Sunday at 15.3%.

The Sunday peak reinforces but does not prove our hypothesis for the weekly distribution of deleted tweets, which will require further study. However, the relative dearth of deleted tweets on Tuesday was unexpected and not readily explainable.

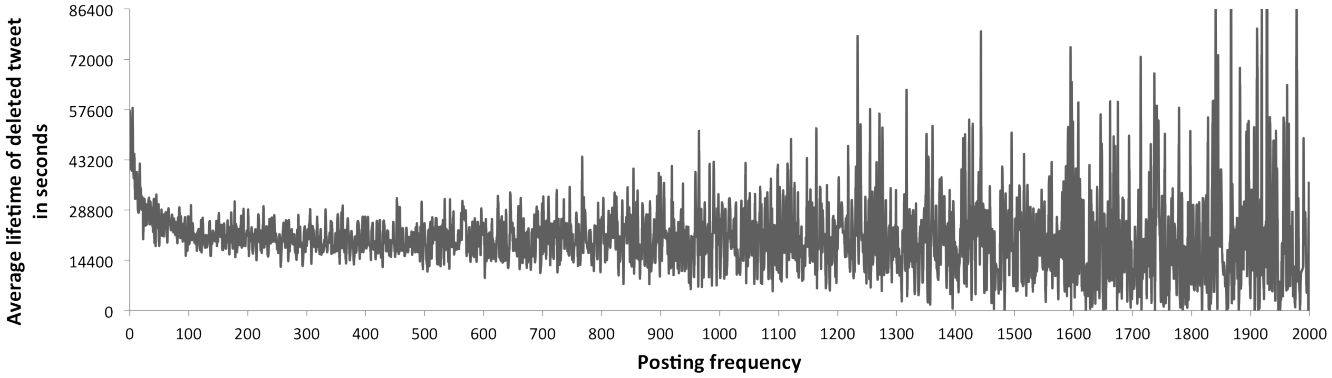


**Figure 9: The relationsip between users' posting frequencies and the average lifespans of their deleted tweets.**

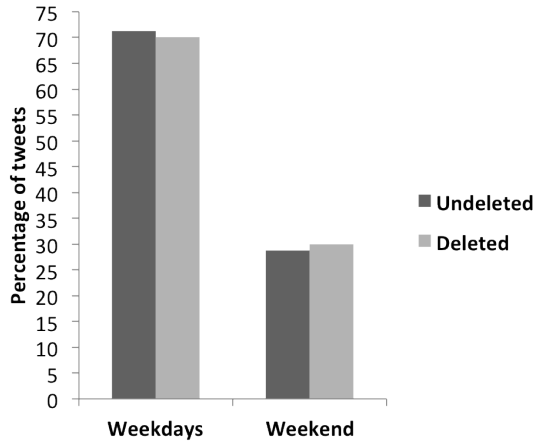


**Figure 10: Distribution of deleted vs. undeleted tweets on weekdays and weekend days.**

### Bulk Deletion

Twitter does not provide a bulk-deletion operation for one's tweets, and some third party applications have been developed for this purpose. Tweeteraser, TwitWipe, Delete Multiple Tweets, Delete My Tweets, and the iPhone applications Delete Your Tweets and Tweeticide are some examples. We attempted to identify bulk deletions in the dataset for two reasons: to determine how widely bulk deletion is used to manage one's Twitter history and to determine the impact of this deletion tactic on our dataset, since it minimizes the element of human intervention in each deletion.

To determine the impact of bulk deletion on the composition of our dataset, we conducted an experiment to determine what fraction of tweets were deleted by bulk deletion services. To do this, we created a chronological record of each user's deletions and experimented with thresholds for time differences between deletions to identify segments of bulk deletion. We assume that the number of tweets in a bulk deletion ($K$) is relatively high, since otherwise the user would delete tweets manually. At the same time, the time elapsed ($T$) from the first event to the last in a bulk deletion set must be very small, since the process is automatic. We experimented with different values of $K$ and $T$, and some of the results are shown in Table 6.

This experiment shows that when $K>5$, the number of bulk-deleted tweets decreases sharply, regardless of the value of $T$. Overall, it appears that a very small percentage of deleted tweets were the result of bulk deletion services. This suggests that bulk deletion has a negligible impact on the results in this paper. Due to the short period of this study, we are not able to make any conclusion regrading whether or not bulk deletion is a common tactic to manage one's Twitter history.

## SPATIAL ANALYSIS OF DELETED TWEETS

Prior research has shown that users care about their location privacy and the granularity at which they disclose their locations in different contexts [21, 16]. In this section we first examine the relationship between location tags and tweet deletion, and then categorize the locations of deleted and undeleted tweets using a service provided by Foursquare.

### Location Sharing on Twitter

Tweeting with location is disabled by default, and users can enable and disable the feature on a per-tweet basis. Twitter's default setting for location granularity is city- or neighborhood-level; however, more fine-grained information sometimes appears in tweets from third party clients. For instance, some third party clients allow a user to share their exact longitude and latitude in a tweet.

There are two ways to delete location information: a user may delete the tweet tagged with the information, or they may delete *all* location information attached to all of their previous tweets, leaving those tweets otherwise unaltered. Twitter does not provide the ability to merely remove the location from a single tweet, which forces users to delete and repost their tweets to accomplish that.

### Deleted vs. Undeleted Tweets with Location

A total number of 494,018 undeleted tweets with locations attached were posted by 12,474 distinct users. For users who posted at least one location, the mean number of tweets posted with location was 40, the median was one, and the standard deviation was 115. However, approximately 50%

| K | T | Instance of Bulk Deletion | % of Deleted Tweets |
|---|---|---|---|
| 5 | 1s | 327 | 0.020 |
| 10 | 1s | 118 | 0.007 |
| 20 | 1s | 37 | 0.002 |
| 5 | 5s | 1265 | 0.078 |
| 10 | 5s | 377 | 0.023 |
| 20 | 5s | 120 | 0.007 |
| 5 | 10s | 2046 | 0.126 |
| 10 | 10s | 57 | 0.035 |
| 20 | 10s | 163 | 0.010 |
| 5 | 15s | 2670 | 0.165 |
| 10 | 15s | 743 | 0.046 |
| 20 | 15s | 197 | 0.012 |

**Table 6: Bulk deletion experiments.**

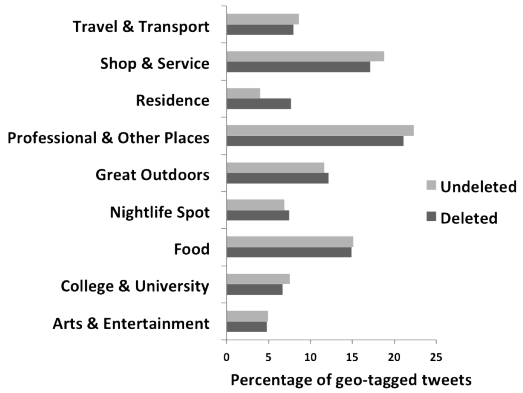


**Figure 11: Distribution of venues' categories of location information attached to deleted and undeleted tweets**

of those users posted eight or less tweets with locations (median = 8). The maximum number of tweets with location per user is 4,454, while the minimum was one tweet per user. The mean number of deleted tweets with location per user was 3.5 tweets, and the maximum was 307. The average ratio of deleted tweets to undeleted tweets with location per user is 3:10. 77.5% of deleted tweets with location were posted from Twitter's official mobile clients (iPhone, Android, Blackberry, Windows, iPad, and mobile web), 6.6% from FourSquare, and 6.4% from TweetCaster. The other 9.5% were posted from a variety of clients including UberSocial (Android and Blackberry) and Instagram.

Although it happened rarely, we observed Twitter users deleting location information without deleting the accompanied tweets.[7] 286 instances of bulk deletion of location information were invoked by 268 users. Although most users (252, or 94%) invoked this action only once, a small number of users (16, or 6%) invoked it more than once: 15 users invoked it twice, and one user did three times.

Tweets that explicitly share exact coordinates of a residence appear in both the deleted and undeleted sets. Such tweets contain text such as "I'm home", "I'm at grandma's house", or "At my friend Alice's house", and they are tagged with exact coordinates as well as other metadata such as the city and state. A manual examination of a random sampling of these coordinates confirmed that they point to residential locations.

### Location Analysis via Foursquare

We hypothesized that deletion rates would vary depending on the locations at which tweets are posted, since some locations are more private or sensitive than others. We were able to test this hypothesis using the Foursquare API, which provides a reverse geocoding service to translate location information (i.e., longitude and latitude) into venues. Foursquare places each venue into one of nine categories: "Arts & Entertainment", "College & University", "Food", "Nightlife Spots", "Great Outdoors", "Professional & Other Places", "Residences", "Shops & Services", and "Travel & Transport".

[7]This data is from a different week (2012-05-16 through 2012-05-22).

Using the reverse geocoding service, it was possible to assign these venue categories to locations associated with 7,604 deleted tweets (89% of all deleted tweets with locations) and 423,673 undeleted tweets (86% of all undeleted tweets with locations).

Figure 11 shows the distribution of venue categories for geotagged deleted and undeleted tweets. Overall, the difference between these distributions is statistically significant ($\chi^2 = 683.5, df = 8$, p-value $<0.05$). More specifically, the percentages of deleted and undeleted tweets in the "Residence" category (respectively 7.7% and 4%) differ significantly ($\chi^2 = 258.5, df = 1$, p-value $<0.05$). The difference between the percentages at "Shop & Service" (17% deleted, 19% undeleted) is also significant ($\chi^2 = 8741.4, df = 1$, p-value $<0.0001$). The other by-category differences were not found to be statistically significant. Notably, these findings agree with those from a prior study of location sharing. Toch et al. [23] determined that residences are particularly private places, where users of location sharing services are more reticent to announce their presence.

## DISCUSSION

Our analysis shows that tweet deletion is a frequent event for a large number of users, and it spans a variety of types of information posted on Twitter. Overall, about 2.4% of all tweets are deleted and about 50% of all users deleted at least one tweet during the weeklong monitoring period. Users deleted tweets that chiefly contained text, tweets with images or locations attached, and location information attached to prior tweets.

The results from the previous sections suggest these answers to the research questions posed in the introduction:

- *The differences between deleted and undeleted tweets are distinct in some dimensions and subtle or nonexistent in others.* Substantial differences appeared in the distributions of clients used to post tweets, the locations of posts (determined via FourSquare), and the quantities of replies that they generated. Slight, inconclusive differences were observed in sentiment vocabulary and how tweets con-

nected to the rest of the Twitter community via mentions and retweets. Substantial differences were expected in the distributions of stereotypically regrettable topics, but these were not observed, perhaps due to the sheer volume and variety of reasons tweets are deleted.

- *A substantial fraction of deleted tweets are deleted for superficial reasons.* Together, typos, rephrasing, and spam account for 18% of deleted tweets. Bulk deletions, however, comprise only a tiny fraction of the sample population. This answer comes with the caveat that small changes in a tweet's textual content (counted as "rephrasings") sometimes may have a dramatic impact on its meaning.
- *It is possible to speculate some common reasons why users delete tweets, but further investigation–including the collection of first-hand explanations of deletions–will be necessary to validate any hypotheses.* Tweets from residences seemed more likely to be deleted than those from many other kinds of locations, possibly reflecting a concern for location privacy. Tweets from web clients were more likely to be deleted than those from other sources, suggesting they are often the source of unwanted tweets that did not meet our criteria for spam.

#### Similarities Between Deleted and Undeleted Tweets

Our analysis shows that deleted and undeleted tweets are largely similar in terms of content, weekday and weekend volume, and volume of tweets with location information attached.

Our lexical analysis of the two sets of tweets revealed that they were fairly similar. For topics that stereotypically might provoke greater deletion, no substantial differences were observed. Although it is likely that some tweets were deleted as a result of containing those topics, more sophisticated methods or further contextual information (beyond what is available on Twitter) will be necessary to discern them. Keyword-based sentiment analysis revealed no differences in positive-sentiment vocabulary, although a slight difference may exist in negative-sentiment vocabulary. Further study will be necessary to determine whether this difference is meaningful. A longer period of data collection combined with more sophisticated topic and sentiment analysis also may produce more substantial results.

Prior to this study, we had informally hypothesized that the percentage of tweets deleted would be significantly higher from mobile devices, due to the challenges of posting (e.g., difficulties with entry methods, cognitive burden from one's surroundings). However, this was not the case, as deletion instead consumed a greater fraction of tweets posted from non-mobile sources. This could be due to the more spontaneous nature of interaction with mobile devices, in contrast with more deliberate interactions with fixed-web devices, on which users may curate their posts with greater ease.

#### Reasons for Deletion

The relative lifespans of spam tweets and typo or rephrasing tweets followed prior intuitions. Posts that needed to be corrected were deleted and replaced relatively quickly, while spam lasted longer, and deleted tweets that satisfied neither of those labels lasted even longer. It is possible that some spam deletions occur because spammers repost their messages (to elevate their positions in followers' timelines), although this was not verified. More advanced methods, and possibly large-scale hand labeling, will be necessary to discern structure in the remaining 82% of tweets that received neither of those two labels.

Although attempts to identify distinct topics among deleted tweets were unsuccessful, our analysis did show slightly more high-frequency negative keywords in deleted tweets than in undeleted tweets. We hypothesize a link between this difference and *regret* as a reason for deletion. Other researchers [25, 5] have studied the causal link between regret and deletion on Facebook, and their results are likely to apply to Twitter as well.

#### Deleting a Tweet Does Not Mean It Is Completely Gone

Deleting a tweet fails to guarantee that it will not be stored elsewhere in repositories outside of the Twitter service. Countless third party applications, such as search engines and clients, access and store users' tweets via Twitter's streaming API. Enforcing deletion across all applications is a challenge, and currently there is no mechanism to assure adherence, even if Twitter demands such compliance. The availability of deleted tweets raises security and privacy implications, both from tweets' textual contents and their location tags. It is incumbent upon Twitter to ensure that users are aware of the limitations of deletion.

Services and entities that archive deleted tweets would find themselves at odds with a proposed reform to the European Commission's data protection rules, which (as of early 2012) includes a *right to be forgotten*: "If an individual no longer wants his personal data to be processed or stored by a data controller, and if there is no legitimate reason for keeping it, the data should be removed from their system" [1]. Although the reform is presently only a draft, it demonstrates the consideration that deletion has received, and that some view the option it represents as a fundamental right.

## CONCLUSION AND FUTURE WORK

To our knowledge, this was the first large-scale study of deleted data from a social network. Our comparisons have illustrated several differences between deleted and undeleted tweets in aggregate, as well as a few unexpected similarities. This study was initially motivated by one of our continuing research goals, to design intervention techniques that "nudge" users to avoid posting tweets that they might regret in the future. The results of this study show that metadata such as location and source client can be pieces of the puzzle to generate effective interventions, while interventions based upon content analysis alone are unlikely to work.

In addition to our intended goal, a few other applications of these findings are possible. Twitter-tailored spelling detection and correction could prevent the necessity of many deletions

and reposts, saving users time and improving their experience. Appropriate training data for this task is readily available in large quantities via the streaming API. Spam filtering by tweet source also appears to be a feasible, albeit with low recall and high precision.

**ACKNOWLEDGMENTS**

This research was supported in part by the National Science Foundation under grants CNS-101276 and CNS-0905562, in part by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office, in part by Google and in part by King Abdulaziz City for Science and Technology.

**REFERENCES**

1. European Commission - Press Release: Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses, 2012-01-24.
2. André, P., Bernstein, M., and Luther, K. Who gives a tweet?: Evaluating microblog content value. In *Proc. CSCW* (2012).
3. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *The Journal of Machine Learning Research* (2003), 993–1022.
4. Bonneau, J. Attack of the zombie photos. *Light Blue Touchpaper* *http://www.lightbluetouchpaper.org/2009/05/20/attack-of-the-zombie-photos/* (2009).
5. Boyd, D., and Marwick, A. E. Social privacy in networked publics: Teens' attitudes, practices, and strategies. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society* (2011).
6. Cheng, A., and Evans, M. Inside Twitter: An in-depth look inside the Twitter world. *Sysomos* *http://www.sysomos.com/insidetwitter* (2009).
7. Cheng, J. Are "deleted" photos really gone from Facebook? Not always. *Ars Technica* *http://arstechnica.com/business/2009/07/are-those-photos-really-deleted-from-facebook-think-twice* (2009).
8. Cheng, J. Three years later, deleting your photos on Facebook now actually works. *Ars Technica* *http://arstechnica.com/business/2012/08/facebook-finally-changes-photo-deletion-policy-after-3-years-of-reporting/* (2012).
9. Hovey, P. Real-time recovery and visualization of deleted tweets. Master's thesis, University of California, 2010.
10. Huberman, B., Romero, D., and Wu, F. Social networks that matter: Twitter under the microscope. *First Monday* (2008).
11. Java, A., Song, X., Finin, T., and Tseng, B. Why we twitter: Understanding microblogging usage and communities. In *Proc. of WebKDD and SNA-KDD* (2007).
12. Krishnamurthy, B., Gill, P., and Arlitt, M. A few chirps about Twitter. In *Proc. WOSN*, ACM (2008).
13. Kwak, H., Lee, C., Park, H., and Moon, S. What is Twitter, a social network or a news media? In *Proc. WWW* (2010).
14. Lampe, C., Ellison, N. B., and Steinfield, C. Changes in use and perception of Facebook. In *Proc. CSCW* (2008).
15. Levenshtein, V. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady* (1966).
16. Lin, J., Benisch, M., Sadeh, N., Niu, J., Hong, J., Lu, B., and Guo, S. A comparative study of location-sharing privacy preferences in the U.S. and China. Tech. rep., Carnegie Mellon University - CyLab, 2012.
17. Naaman, M., Boase, J., and Lai, C. Is it really about me?: Message content in social awareness streams. In *Proc. CSCW* (2010).
18. Nielsen, F. Å. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *ArXiv e-prints - arXiv:1103.2903* (2011).
19. Ramage, D., Dumais, S., and Liebling, D. Characterizing microblogs with topic models. In *Proc. ICWSM* (2010).
20. Raynes-Goldie, K. Aliases, creeping, and wall cleaning: Understanding privacy in the age of Facebook. *First Monday* (2010).
21. Sadeh, N., Hong, J., Cranor, L., Fette, I., Kelley, P., Prabaker, M., and Rao, J. Understanding and capturing people's privacy policies in a mobile social networking application. *Personal Ubiquitous Computing* (2009).
22. Sysomos. Twitter Statistics for 2010: An in-depth report at Twitter's growth 2010, compared with 2009. *http://www.sysomos.com/insidetwitter/twitter-stats-2010/* (2010).
23. Toch, E., Cranshaw, J., Drielsma, P. H., Tsai, J. Y., Kelley, P. G., Springfield, J., Cranor, L., Hong, J., and Sadeh, N. Empirical models of privacy in location sharing. In *Proc. Ubicomp* (2010).
24. Tufekci, Z. Facebook, youth and privacy in networked publics. In *Proc. ICWSM* (2008).
25. Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., and Cranor, L. F. "I regretted the minute I pressed share": A qualitative study of regrets on Facebook. In *Proc. SOUPS* (2011).
26. Wu, S., Hofman, J., Mason, W., and Watts, D. Who says what to whom on Twitter. In *Proc. WWW* (2011).
27. Zhao, D., and Rosson, M. How and why people Twitter: The role that micro-blogging plays in informal communication at work. In *Proc. Group* (2009).