# A Large-Scale Analysis of the Twitter Social Network

**Daniel L. Schafer**

May 2010

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**Senior Thesis Advisor:**
Prof. Luis von Ahn

*Submitted in partial fulfillment of the requirements
for a Senior Thesis*

## Abstract

Social networking websites have become increasingly important tools for communication and interaction, and have collected data sets of remarkable size. Over a period of several months, a large corpus of data was collected by crawling the Twitter social networking website. The corresponding analysis examines over 65 million accounts, over two billion messages, and more than two billion directed connections between users. The network characteristics of the social network and the user behavior on this network were studied. Significant shifts from the expected behavior of such a system were found, suggesting the existence of anomalous accounts, and potential methods to detect such users are suggested. The centrality of users in the social network was determined, with the effects of user properties on this centrality examined. A website was created containing the aggregated information from the data set and the results of the aforementioned analyses.

# Contents

## List of Figures

## List of Tables

# 1 Introduction

This thesis is divided into two major parts. The first part of the thesis describes the collection and analysis of a large data set from the Twitter social networking website. More than 66 million users, two billion social connections and two billion posts were gathered, and the properties of this graph were analyzed. Anomalies in the data set were identified, and the temporal posting tendencies of users was studied. Additionally, the effect of posting tendencies on the centrality of nodes in the network was analyzed. This part of the thesis is described in more detail in Section 2.

The second part of the thesis discusses the creation of a website to allow easy access to the information in the data set and results of the analysis. One component of this site displays the characteristics of Twitter users, including the results of analyses currently unavailable elsewhere on the internet. Another page enables users to track the usage of phrases on Twitter over time, and will eventually display explanations for why spikes and dips in the graph occur. A third page identifies the most important users of Twitter according to our analysis. This website was christened TweetMine, and its creation and potential impact is described in Section 3.

# 2 Twitter Data Set

## 2.1 Fundamentals of Twitter

Twitter is a short messaging service that allows users to publish *tweets*: messages no longer than 140 characters [1]. This subsection is designed to introduce the reader to the basic structure of Twitter. It will also cover some of the traditions among Twitter users, and how these traditions provide additional metadata about tweets. Throughout this subsection, we will use dlschafer and bmeeder as example Twitter users (these are the actual Twitter usernames of the author and a collaborator).

### 2.1.1 Signing Up and First Tweets

A user interested in Twitter begins by signing up for a Twitter account. The initial signup page asks the user to provide their full name, a username, a password and their e-mail address, and contains a reCAPTCHA to prevent automated registrations. Once the user has set up their account, he will be issued a *Twitter profile page* based on their username; for example, the page for user dlschafer is `http://twitter.com/dlschafer`. Now that the user has an account, he can begin publishing status updates, or tweets, which will be visible on the user's profile page. Additionally, when the user is logged in to Twitter, Twitter will customize the *Twitter home page*, which the user views at `http://twitter.com`.

By default, the new user's tweets will be visible to anyone; even visitors without accounts can see them simply by visiting the profile page for that user. However, it is possible for a user to mark their account as *protected*, in which case only authorized visitors will be allowed to see the user's statuses. For example, if dlschafer were to protect his account, an unauthorized viewer of the profile page for dlschafer would not see his tweets, but would instead see a notice reading "This person has protected their tweets."

### 2.1.2 Social Graph on Twitter

Once a user has set up his account on Twitter, he is encouraged to *follow* other users. By doing so, the followed user's tweets will appear on the following user's Twitter home page, enabling the following user to keep up to date on the followed user's activity. For example, since dlschafer follows bmeeder, anytime bmeeder posts a Twitter update, it appears on dlschafer's Twitter home page. Following is not necessarily symmetric; dlschafer can follow bmeeder without bmeeder following dlschafer back. If dlschafer is following bmeeder, then dlschafer is referred to as a *follower* of bmeeder, and bmeeder is referred to as a *friend* of dlschafer.

This creates a social network with directed edges among twitter users; when dlschafer follows bmeeder, we create a directed edge from dlschafer to bmeeder, creating a directed graph. The direction of this edge is somewhat arbitrary; we could have instead directed the edge from bmeeder to dlschafer in our graph. This direction was chosen for convenience in running later analyses.

For a protected user account, visibility of tweets is restricted only to that user's followers, and follower requests must be confirmed by the account owner. For example, if dlschafer's account is protected, bmeeder will initially be unable to see dlschafer's tweets, getting the protection notice described previously when he visits dlschafer's profile page. If bmeeder follows dlschafer, dlschafer will receive a *follower request*. Only after dlschafer confirms the follower request will his profile page become visible to bmeeder; at this point, dlschafer's tweets will also begin appearing on bmeeder's home page.

### 2.1.3 Message Conventions

**@-mentioning** A Twitter user can *@-mention* another user in a tweet by typing the other user's username preceded by the @ symbol. For example, dlschafer might tweet "I am working on my research with @bmeeder",@-mentioning bmeeder in his tweet. This affects the readers and users of the site in a few ways. First, bmeeder will receive a notification that he was mentioned in a tweet. Additionally, when this Tweet is displayed on the website, the "@bmeeder" portion of it will contain of a link to bmeeder's profile page, allowing readers of the tweet to follow this link and learn more about bmeeder.

One specific use of @-mentioning is *@-replying*, where a user will begin a tweet with an @-mention. For example, in response to the dlschafer tweet above, bmeeder might tweet "@dlschafer We are being so productive right now!". In addition to having all the properties of @-mentions, @-replies associate the tweet as a reply to the specified user. In certain circumstances, the metadata on the tweet might have even more information. If bmeeder created that tweet by clicking the reply option on dlschafer's original tweet, then Twitter will note that this tweet was a reply to a particular post, and store this along with the tweet. In this case, viewing the tweet on the website will indicate this metadata by noting the tweet was "in reply to dlschafer", where the "dlschafer" portion links to dlschafer's original tweet. This type of @-reply is called an *@-response*.

**Retweets** Often, a Twitter user will see another user's tweet and wish to rebroadcast that message. This was traditionally accomplished with a *retweet*. A user will preface their tweet with the token "RT", @-mention the user who originally posted it, and include the original message. For example, if dlschafer tweeted "It is sunny in Pittsburgh." and bmeeder wanted to retweet that message, he would tweet "RT @dlschafer It is sunny in Pittsburgh."

At the time the data set was gathered, retweets were simply a tradition among Twitter users; there was no built-in functionality of the official Twitter site to facilitate the creation of retweets, or recognize that a given tweet is a retweet. Since then, Twitter has formalized this aspect of Twitter communication, and allows for applications to request all retweets of a given post [3] [4].

**Shortened links** One major use of Twitter is for sharing links of interest, however, the 140-character limit of tweets makes it difficult to share long links. Hence, when sharing links, Twitter users tend to use URL shorteners: services that create a shorter URL that links to the original one; Twitter also offers a built-in URL shortener, so if the user enters a long URL, it will be replaced by the shortened one in the tweet. Twitter used to use TinyURL for this service, but switched in May of 2009 to bit.ly, a competing service [15]. As an example, bit.ly shortened the URL `http://www.cs.cmu.edu/~bmeeder/` to `http://bit.ly/zI9Ur`.

One interesting aspect of bit.ly is that when issued the same link to shorten multiple times, it will issue different shortened URLs in each instance. Hence, the propagation of a given message containing a URL can be measured by tracking the bit.ly link contained therein, as that URL should be unique to that message and retweets of that message.

**Hash-tags** Twitter users often wish to tag their statuses with an identifier, either to make it easier to search for or to otherwise identify that tweet as being related to some external item. Twitter enables this through the use of *hash-tags*, where a status will include a topic identifier preceded by the octothorpe symbol (#). These hash-tags might appear at the end of a post, or they could be embedded in the middle. As an example, in the months before this thesis was submitted, dlschafer might have had the status "Working on his #seniorthesis submission", or alternately, "Working on his paper for the next hour. #seniorthesis". Both of those tweets would have appeared if another user searched for the #seniorthesis hash-tag.

**Table 1:** Summary of Collected Data

| | |
|---|---|
| **User profiles:** | 66,250,639 |
| **Messages collected:** | 2,022,696,632 |
| **Network edges:** | 2,032,612,302 |

One example of hash-tag usage on Twitter is #FollowFriday; users will list other people they think are interesting users, then tag that post with #FollowFriday or #ff. Users looking for interesting people to follow can then search for that hash-tag, and get suggestions for who to follow from other's posts. This hash-tag, unsurprisingly, displays cyclic popularity depending on the day of the week. Another common hash-tag is #nowplaying, which users place in front of the music they are currently listening to. Unlike #FollowFriday, this hash-tag displays relatively constant usage over time. Other hash-tags spike in usage around specific events; in the aftermath of the Iran Election in 2009, #IranElection was popular, while during the 2010 Super Bowl, #Colts, #Saints and #SuperBowl gained in popularity.

## 2.2 Measurement Methodology

We have collected the content of the Twitter social network using Twitter's publicly available Application Programming Interface (API). Twitter implements the API using HTTP methods that accept or return data in a structured format such as XML or strings in JavaScript Object Notation (JSON) form. To avoid excessive use and abuse of this service, the number of requests per client is limited; the baseline number of requests is 150 per hour per client. However, users can request their screen name or IP address to be *white-listed*, at which point they can make 20,000 requests per hour from that screen name or IP address. A summary of the quantity of data we have collected is found in Table 1.

### 2.2.1 Measured Features

The Twitter API provides access to all of the information users can see when visiting the website, along with other information that isn't normally available through the web interface. We primarily use four sets of API methods to get information for each user. One method is used to retrieve information about a user, two methods are used to acquire social network connections, and a fourth method is used to access messages that have been generated. We summarize the information provided by each in Table 2.

Information from the users/show API call is always available, regardless of whether a user's account is protected or not. In particular, the friend, follower and status count are always available, and the user optional fields will be available assuming the user has filled them out. Information on who a user is following, who follows that user, and the contents of a user's posts are only available for unprotected accounts. For statuses posted by unprotected accounts, all information about that status is made available. In particular,

**Table 2:** Summary of API methods used

| API Method | Retrieved Information |
|---|---|
| **users/show**<br>(always specified) | user ID, screen name, account creation time, number of friends, followers, tweets, favorites |
| **users/show**<br>(user optional) | ‘real’ name, location, description website, time zone |
| **followers/show** | IDs of users following specified user |
| **friends/show** | IDs of users followed by specified user |
| **statuses/show** | unique message ID, post content, posting user ID, in response to user ID, in response to message ID, post time, update method |

the in-response-to-post and in-response-to-user fields specify whether a particular message is a reply to a specific message, or is a message directed at an individual. However, it does not seem as though this feature is implemented uniformly across all methods of posting messages to Twitter; applications may not include this metadata when they upload posts to Twitter, for example.

### 2.2.2 Crawl Infrastructure

Because Twitter limits the number of requests to its API to 20,000 requests per hour, it is possible to collect data at the scale we have only by splitting the queries between a large number of machines. We hence created a distributed infrastructure using a cluster of 80 machines, and got two Twitter usernames white-listed for API queries. This means each of the machines can issue 20,000 unauthenticated requests per hour, plus an additional 20,000 for each of the white-listed users. This infrastructure allows us to crawl twitter at 240 times faster than with a single machine, meaning a full user crawl can be done in around two days.

### 2.2.3 Limitations of the Collected Data

We believe that the data we have collected is extremely comprehensive and is the largest such collection to date; however, some information about the network is missing. In particular, certain limitations of the current API prevent us from accessing all of the information we would like. For example, only the last 3,200 messages posted by a user are made available. This means that we have an incomplete view of a user who has generated more 3,200 messages by the time we first crawl them. Additionally, all links between users are publicly available but Twitter does not include the time at which the links were created. In order

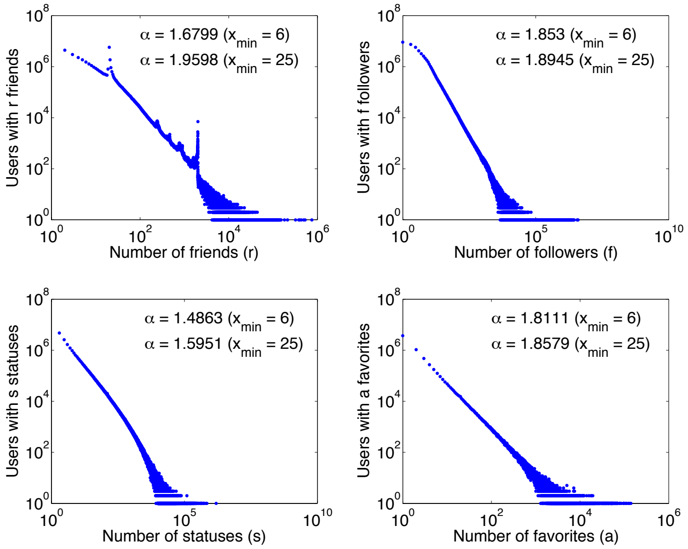


**Figure 1:** Log-log plot of number of users by friend, follower, tweet and favorite count

to study the growth and evolution of the network structure, we must repeatedly query follower and following information.

Despite this limitations, we believe that the data we have collected allows us to thoroughly analyze many aspects of Twitter. Even though we couldn't get a complete message history for some users, these users have generated a total of only 400 million messages that could not be downloaded. We can get around the limitation of not having creation and deletion times of social links by repeatedly crawling users. Using a prioritization scheme based on the number of friends and followers a user has, as well as how recently the user has joined the network, we can get snapshots of the social network at a resolution of one day or finer.

## 2.3 Examining the Twitter User Base

### 2.3.1 Analysis of the Twitter Social Graph

One of the most basic aspects of the social graph is its degree distribution. As the follower relationship on Twitter induces a directed graph structure, we analyze both the in-degree and out-degree distributions. A log-log plot of the degree frequency for friends and followers appear as the first two plots in Figure 1. We see a clear power law relationship here, which agrees with previous analyses of social network degree distributions in [13] and [7]. We computed the power-law coefficient for these graphs using the method described in [8]. We examine the results for both $x_{min} = 6$ (the recommended $x_{min}$ from that paper) and $x_{min} = 25$, since there is an anomaly in the friends graph at $x = 20$. We find that the $\alpha$

**Table 3:** Power-law $\alpha$ values

| **Graph** | $\alpha$ | |
|---|---|---|
| | $\mathbf{x_{min}} = 6$ | $\mathbf{x_{min}} = 25$ |
| Friend | 1.6799 | 1.9598 |
| Follower | 1.853 | 1.8945 |
| Status | 1.4863 | 1.5951 |
| Favorite | 1.8111 | 1.8579 |

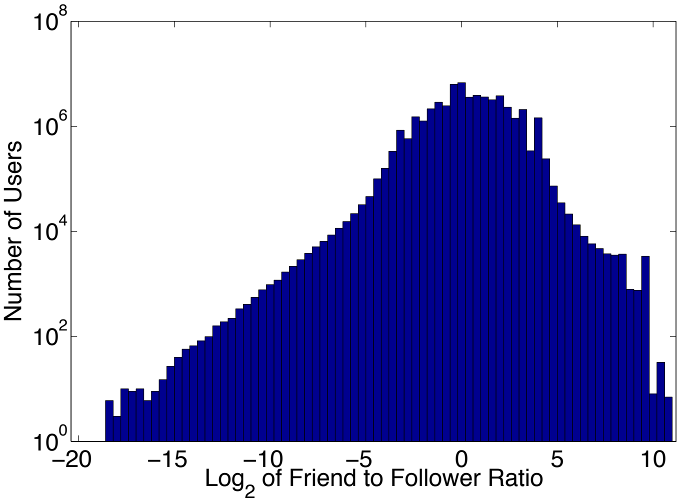


**Figure 2:** Ratio of friends to followers

value does not change dramatically for followers, statuses, or favorites between these two values, but that for friends, the $\alpha$ dramatically increases when we set $x_{min}$ to avoid the anomaly at $x = 20$. We find that the $x_{min} = 25$ coefficients have $\alpha$ slightly less than 2 for both friends and followers; this differs from the 2007 figure of $\alpha \approx 2.4$ found in [10]. Table 3 contains all of the power-law $\alpha$ values.

It is interesting to observe how the number of friends and followers is related for a given user. For example, celebrities tend to have many more followers than they do friends, while an account that rarely posts but wants to aggregate content would have more friends than followers. The log-log distribution of friend to follower ratio is plotted in Figure 2, for those users for which it is well defined; in particular, users with a 0 ratio (for they have no friends) or a $\infty$ ratio (for they have no followers) are omitted from this graph.

Unsurprisingly, we note the peak is near 0, where users have a similar number of followers as friends. However, we note that the left side of the graph (more followers than friends) has a much shallower drop-off than the right side of the graph (more friends that followers). There are a few reasons for this. First, a low friend-to-follower ratio is typical

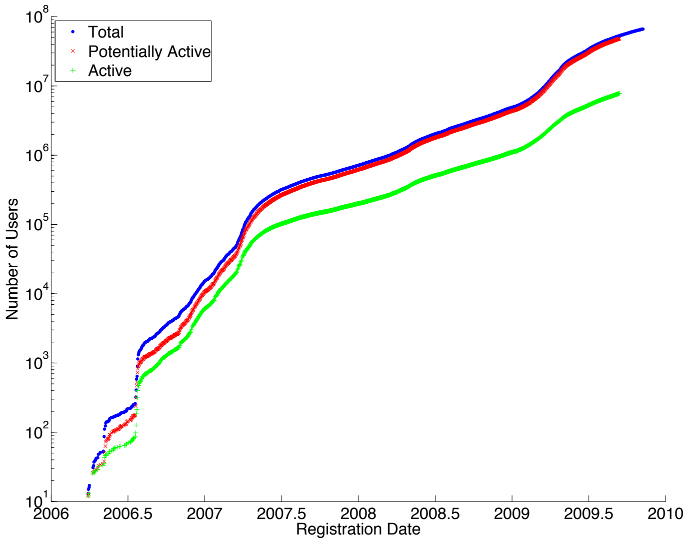


**Figure 3:** Cumulative distribution of Twitter users by registration date

of a celebrity account on Twitter, many of which have millions of followers, but only a few hundred friends. Because of this celebrity effect, it is possible for a significant number of accounts to have extremely large numbers of followers with few friends; prominent public figures can attain thousands of followers quite quickly.

On the other hand, it is fairly difficult to gain a large number of friends without having any followers. First, many accounts on Twitter *follow-back*, where they will follow anyone who follows them as a form of courtesy; in gaining 100 friends, it is likely a user will gain at least a few followers. Additionally, to slow spammers, Twitter at times places restrictions on the number of friends a user can have; this might prevent accounts from gathering large friend counts with low follower counts.

### 2.3.2 Understanding User Lifespan

The collected user data includes registration times, which allows us to plot the growth of Twitter over time. The top plot in Figure 3 plots the total number of accounts on Twitter over time, with the $y$-axis on a logarithmic scale. However, not every registered user remains an active user of the service. For our discussion, we will define an *active user* to be one who has posted a Twitter status update in the last month.

To determine if an account is active or not, we will examine the latest-status data provided by the API for that user. As this data is unavailable for protected accounts, all discussion of active users will be restricted to unprotected accounts. Additionally, we will restrict discussion of activeness to users who registered more than two months ago; including brand-new users in a discussion of activeness will naturally skew the results, as

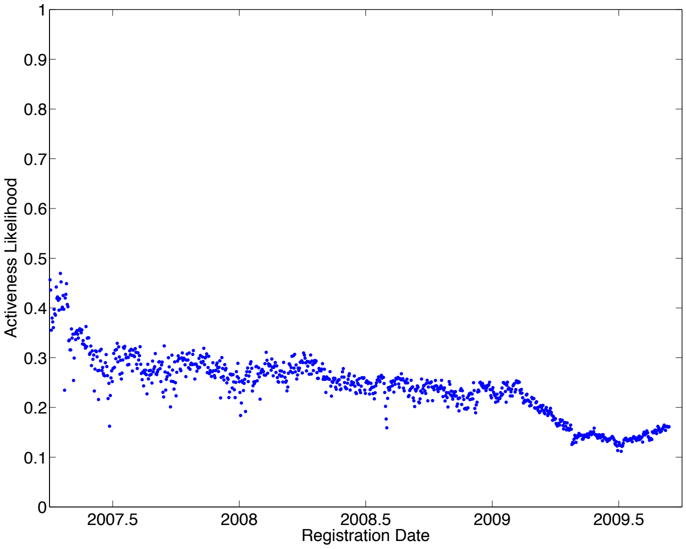


**Figure 4:** Activeness Likelihood vs. Account Creation Date

any user who registered recently has definitionally interacted with the site recently. We will refer to users who are at least two months old and have unprotected accounts as *potentially active users*, and we will define the *activeness likelihood* of a set to be the number of active accounts in that set divided by the number of potentially active accounts in that set. Figure 3 contains two additional plots; the number of users who remain potentially active today, and the number of users who are active today, both plotted against their registration time. Note that these plots cut off two months prior to present day due to our definition of potentially active and active users. All three plots in this figure have been experiencing exponential growth. However, the number of users active today stays consistently below the number of potentially active users today; in particular, 1/6 of the potentially active users are actually active at our last measured point. Projecting forward two months, it seems likely, then, that of our total count of around 66 million users registered, around 11 million of those could be considered active.

Based on those two plots, we can plot the *activeness likelihood* against the registration time of a user. The data is messy due to sparsity for the very early days of Twitter, so Figure 4 begins in the middle of 2007. The activeness likelihood seems to hover around 0.25 (that is, 25 percent of users registering on a given date remain active today), though it begins dipping closer to 0.15 as we approach the present time. From this, we can conclude that initial adopters of Twitter (those who joined before 2009) are more likely to remain active today, and that those who joined in the last year are less likely to remain active users of the site. The activeness likelihood then rises as we get closer to present day. It is unclear whether this is because more user activeness is actually on the rise again among

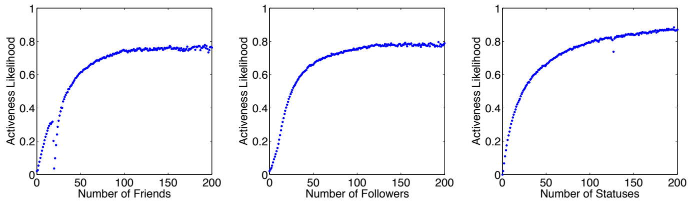


**Figure 5:** Activeness Likelihood vs. Friends (left), Followers (center), and Statuses (right)

newer users, or whether this is simply the result of their registration time being closer to our threshold for activeness, and hence the initial engagement that comes with registration is altering the activeness data.

Finally, we would anticipate that as a user's friend, follower and status count increases, that user would become more likely to remain active. Figure 5 plots the activeness likelihood against the number of friends, followers and statuses a user has. As expected, these graphs are almost entirely monotonically increasing, and they seem to approach around .8, which suggests that among users who were at one point extremely invested in Twitter (having attained 200 friends, 200 followers, or 200 statuses), around 80 percent remain active today. There are two anomalies in these graphs; at around 20 friends, where activeness dips to below five percent, and at exactly 127 statuses, where activeness drops by around 10 percent. This is partially due to anomalous accounts, and is discussed further in Subsection 2.5.

### 2.3.3 Account Age and Protected Status

One interesting property of Twitter users that has changed over time is their tendency to protect their accounts. Figure 6 plots the percent of users who protect their account against registration time. A dramatic drop in protection percentage is immediately apparent from this graph; this shift occurred on April 21, 2007. We can find no specific incident that might have led to this change; it is possible that Twitter may have altered their registration flow at this time to make the ability to protect one's account slightly less visible. We also note that this shift aside, more recently registered users tend to be significantly less likely to protect their accounts, possibly because of a shift in the usage patterns of Twitter users as the site grew in size.

### 2.3.4 Geographic Composition

One interesting aspect of Twitter is where users are tweeting from. Twitter allows users to list their time zone; though it does not require it, and around 38 percent of users fill this field out. However, the 62 percent of users who leave the field blank account for only 11 percent of all statuses! This could be because users invested in Twitter are more likely to fill out a complete profile. Furthermore, the tweets from someone who omitted that

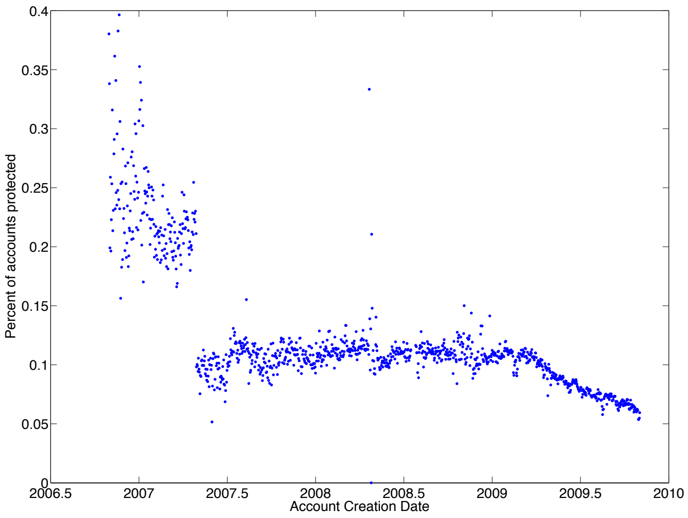


**Figure 6:** Percent of accounts protected by registration time

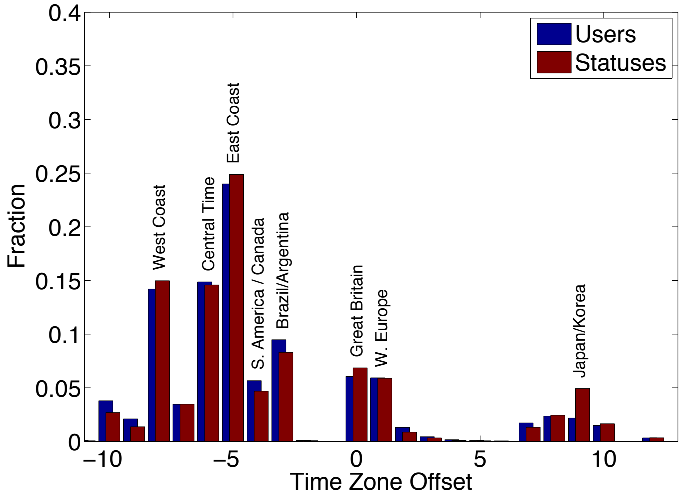


**Figure 7:** Percentage of users/tweets from each time zone

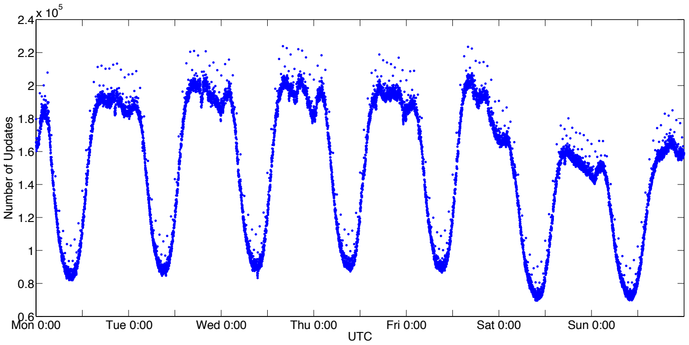


**Figure 8:** Number of tweets each minute of week

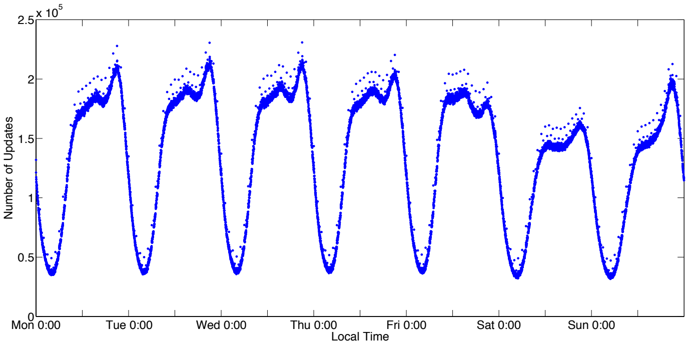


**Figure 9:** Number of tweets each minute of week (Localized)

field will have inaccurate timestamps, which might prompt a user with a lot of tweets to update their profile data to fix this. For the remainder of this section, users lacking time zone information are omitted.

Figure 7 shows the number of tweets and users from each time zone, ranging from UTC-11 to UTC+13. A plurality of users are on the east coast of the United States, followed by the West Coast and Central time zone of the United States. Most other locales are significantly lower, though the time zones corresponding to Brazil and Argentina, Great Britain, and Western Europe have significant user bases. Notably, UTC+9 (corresponding to Japan and Korea) has a significantly higher percent of tweets given their number of users; despite only having 2% of Twitter's user base, they contribute 5% of Twitter's posts. It is not currently known why this occurs.

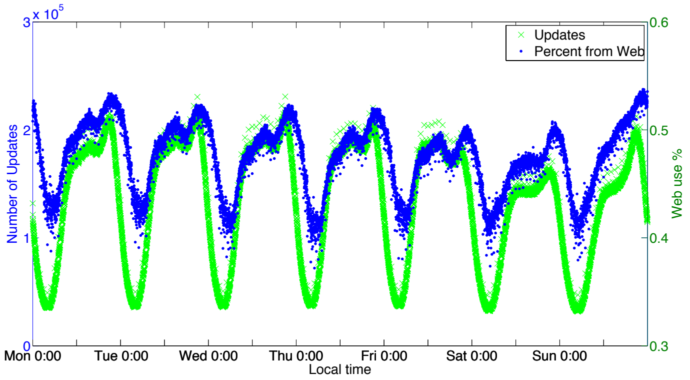


**Figure 10:** Percent of tweets from web each minute of week (Localized)

## 2.4 Analysis of Twitter Messages

### 2.4.1 Timing of Tweets

With our record of all public tweets, we were able to determine how tweet frequency changes over the course of a week. A plot of the number of tweets by minute appears in Figure 8. This plot has absolute time along the $x$-axis; a user in London posting at midnight and a user in New York posting at 5:00 AM appear as posting at the same time. While this allows us to observe the times when Twitter is seeing the most traffic, to observe the behavior of Twitter users, we would like to correct for time zone offset. Figure 9 shows the same plot as before, but this time corrected for time zones (so all data points for midnight are now for users who posted at midnight in their local time zone).

We note that the peak usage time for Twitter on weekdays is at 9:00 PM. This is notably different than the results of [12], which found a main peak at 3:00 PM, and two smaller peaks at 10:00 AM and 9:00 PM. Both tweets from the website and tweets from non-web sources showed essentially identical general shapes to the overall graph, and both web and non-web statuses reach their peak at 9:00 PM.

We also note that the graphs are essentially the same for Monday through Thursday, but that the weekend displays entirely different behavior. Posting is significantly reduced on Friday nights, so much so that 9:00 PM is no longer the peak on this day alone. Saturday also shows reduced posting, though it appears more like a miniaturized version of a weekday. Sunday starts slowly, but the 9:00 PM peak returns in full weekday-like force by Sunday evening.

The percent of tweets coming from the website (rather than applications) over time is shown in Figure 10. Note that web traffic percentage follows the overall number of tweets quite closely; times where tweeting is most frequent shows the most use of the website. This suggests that Twitter observes a more constant flow of application based tweets, whereas the website traffic varies more with time.

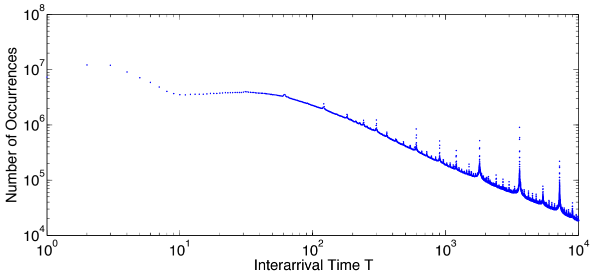


**Figure 11:** Distribution of Tweet Interarrival Times

One noteworthy aspect of these plots is the increased activity on the minute corresponding to the top of each hour (these are the single outlier points appearing 10% higher than the rest of the graph, most noticeable in Figure 9). This is analyzed more closely in Subsection 2.5.

### 2.4.2 Time Gaps between Tweets

One way we can analyze the engagement of Twitter users is by determining how long of a gap users have between consecutive tweets; large gaps mean the user is not using Twitter for large periods of times, whereas small gaps indicate users are consistently engaged with the service. Hence, for each user, we can determine their average gap between tweets; a histogram of this value appears as Figure 11, with a logarithmic axis for time.

Note again that in this plot, we note a large number of users with average tweet gap times near round numbers; an abnormally large number of accounts have an average tweet gap of exactly one hour, for example. This, too, will be analyzed in Subsection 2.5.

## 2.5 Anomalous Accounts in the Data Set

During the course of our data analysis, we found significant evidence for the existence of automated or spamming accounts. Social networking websites must actively combat spam to remain effective, and Twitter is particularly appealing to spammers because of the popularity of trending topics and the large user base of the website. In this section, we do not attempt to distinguish between accounts that are run by spammers and legitimate accounts that automatically generate content. Rather, we focus on anomalies in the aggregated data to estimate the prevalence of automated accounts in Twitter. The discoveries noted below are not proven to be caused by anomalous accounts; instead, they merely demonstrate the likely existence of such accounts, and suggest potential future work on discovering those accounts.

### 2.5.1 Regular Message Timing

Examining message creation time provides the first indication that a sizable number of messages are generated automatically. In particular, one can examine the minute of the hour in which messages are posted, collected over the entirety of our message data; this was Figure 9. Immediately, one notices that in the first minute of every hour there is approximately a ten percent increase in traffic for that minute; that is, ten percent more tweets are posted at 10:00 PM as compared to 9:59 PM and 10:01 PM. We hypothesize that this increase of ten percent is due to automated accounts; it seems unlikely that human users would consistently post exactly on the hour. Additionally, there is a noticeable drop in the percent of posts coming from the web each hour, so it is likely that the anomalous posts on the hour are coming from non-web sources, which are likely easier to automate.

We also measured the time between messages for each user, and aggregate these time deltas over all users; this was Figure 11. We find again that the plot is not smooth around gaps of very regular sizes: one hour, two hours and one day, for example. It is likely that this is also the result of automated accounts posting on regular intervals.

### 2.5.2 Social Graph Anomalies

Looking at the vertex degree distribution, it is immediately apparent that the friends distribution is not well behaved. In particular, there is an enormous jump in the number of people following approximately twenty people, and again in the number of people following approximately two thousand accounts. The anomaly at following two thousand people might be a result of Twitter imposed limits created to reduce the amount of 'spam following' in the network.

A feature of Twitter registration can also explain the large number of users following approximately twenty people. We posit that this number results from the fact that the Twitter web interface offers users an initial selection of twenty users they can follow. Hence, rather than starting from 0 users, users might start from 20 users instead. This is supported by the low activeness among users with 20 friends; 20-friend users should and do behave more like 0-friend users than 19 or 21-friend users. Hence, the spike in users and drop in activeness at 20 friends can probably not be used effectively to detect anomalous accounts.

We also noted that users with exactly 127 statuses are significantly less likely to be active, as seen in Figure 5. It is possible that an automated account will post exactly 127 statuses to look more like real users, then will cease posting, making them inactive. Hand-investigating accounts with 127 statuses revealed many of them are anomalous; they will often post exactly 127 statuses all using the same hash-tag, perhaps in an attempt to move that hash-tag into the trending topics. The choice of 127, however, is peculiar; it is the maximum value of a signed byte, but why this is significant to the operators of these accounts is unknown.

| In-degree | | PageRank (scaled by 10,000) | |
|---|---|---|---|
| User | In-degree | User | PageRank |
| aplusk | 2,968,120 | BarackObama | 18.2497 |
| britneyspears | 2,900,190 | cnnbrk | 15.0292 |
| TheEllenShow | 2,877,175 | twitter | 10.4637 |
| cnnbrk | 2,812,327 | nytimes | 7.6198 |
| BarackObama | 2,770,942 | KimKardashian | 7.2160 |
| twitter | 2,701,888 | levarburton | 6.7163 |
| KimKardashian | 2,644,440 | shitmydadsays | 6.4186 |
| nytimes | 2,135,172 | ev | 6.1158 |
| RyanSeacrest | 2,057,453 | TheEllenShow | 5.9818 |
| johncmayer | 1,987,951 | aplusk | 5.7939 |

**Table 4:** Top users for in-degree and PageRank centrality on Twitter

## 2.6 Centrality on Twitter

A natural analysis to perform on that Twitter data set is a centrality analysis, to determine which nodes in the network are potentially influential. We analyzed the in-degree centrality [9] of the Twitter network, as well as finding the PageRank of each node in the network[14].

### 2.6.1 Degree Centrality

Degree centrality is an easy measure to compute on Twitter; when data is gathered on a given user, their in-degree is provided in the form of their follower count. Hence, finding the top 10 accounts for in-degree centrality simply involved sorting on that field, and no special tools were required. Those 10 users (and their follower counts at the time of data collection) appear in Table 4.

### 2.6.2 PageRank Centrality

Computing the PageRank on the gathered data set was significantly more difficult computationally than computing the in-degree centrality. To perform this analysis, the PEGASUS system was used[11]. Using the functionality of this system, the top 10 users in PageRank were determined, and appear in Table 4.

### 2.6.3 Predicting PageRank from local properties

While determining the PageRank of a user given the entire graph is a matter of computation, one might wonder if any of the local user properties correlate with high PageRank. To visualize the correlation, various user properties were plotted against the average PageR-

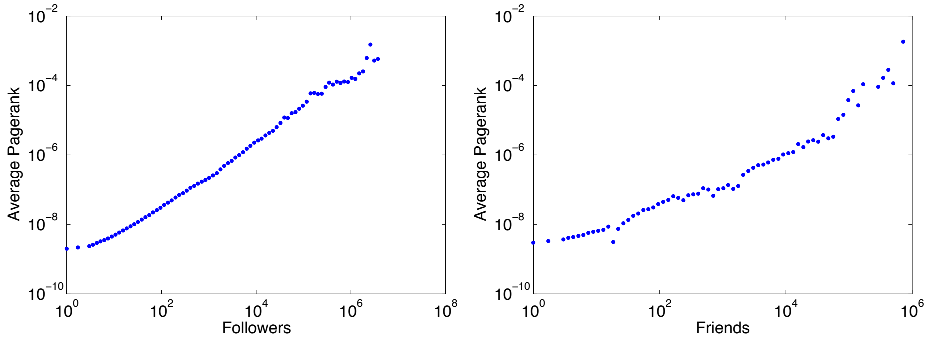


**Figure 12:** Average PageRank for users with varying degrees

ank of users with that property. To avoid data noise, similar users were clustered together; for example, users were clustered based on the log of their degrees in the first set of graphs.

**Degree** Because of the design of PageRank, one gains PageRank by acquiring in-links from other high PageRank users. Hence, we would expect that users with higher in-degrees would have correspondingly higher PageRanks; this pattern was unsurprisingly observed. The correlation of out-degree with PageRank is more surprising; one possible explanation is that the tradition of "following-back" on Twitter leads users with a large in-degree to also have a large out-degree, causing this correlation.

The graphs show a nearly straight-line correspondence between in-degree and PageRank; as these are log-log plots, this suggests that the average PageRank of users with in-degree $D$ is equal to $D^c$ for some constant $c$, where $c$ appears to be around 1. The plots of in-degree and out-degree vs. average PageRank are in Figure 12.

**Posting Frequency and Account Age** Posting frequency turned out to be heavily correlated with PageRank as well. When only the raw number of statuses is considered, the average PageRank for a given number of statuses rises until around 10,000 statuses, at which point it begins decreasing. This is logical, as accounts with too few statuses are not likely to be interesting, and too many statuses is likely to annoy users with constant posting.

To account for varying account age, we plotted average PageRank against the average gap between posts. The overall shape of this graph is similar, with a peak at one post every three hours. However, there are two smaller peaks, once when the gap is extremely large and once when the peak is extremely small. No explanation for these smaller peaks is known at this time, though it should be noted that this gap is measured over the course of the account's lifetime, a user with a small gap might be a user who registered, posted once, then left the site.

Account age has a similarly strong effect on PageRank. Older accounts, on average,

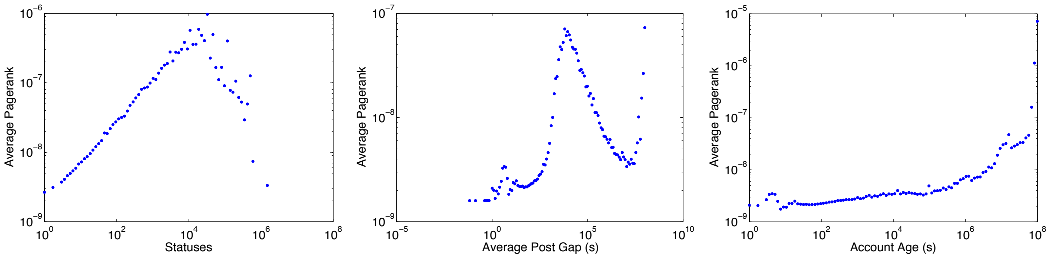


**Figure 13:** Average PageRank for users with varying posting frequency and account ages

**Hash-tag** The ratio of total hash-tags used to total statuses.

**Mentions** The ratio of total mentions used to total statuses.

**Replies** The percent of statuses that are replies to a user.

**Responses** The percent of statuses that are responses to a post.

**Retweets** The percent of statuses that are retweets.

URL The ratio of total URLs used to total statuses.

**Table 5:** Post Measures for Users

have higher PageRanks, especially once particularly old accounts are considered. The plots of statuses, post frequency and age vs. average PageRank are in Figure 13.

**Post Contents** For each user, we analyzed the contents of their public posts in aggregate. By doing so, we came up with six measures for each user, contained in Table 5. However, plotting these measures against average PageRank yielded almost no correlation. The plots of these measures vs. average PageRank are in Figure 14.

# 3 TweetMine Website

## 3.1 General Idea

The analysis of the Twitter data set was done in an academic setting, but the results might also be of interest to the general public. A website was thus created to display the information gathered and analysis performed. This website was named TweetMine, and should be launched in May 2010.

For notational convenience, Twitter accounts will continue to be called “users,” while the person viewing the site will be called a “visitor.”

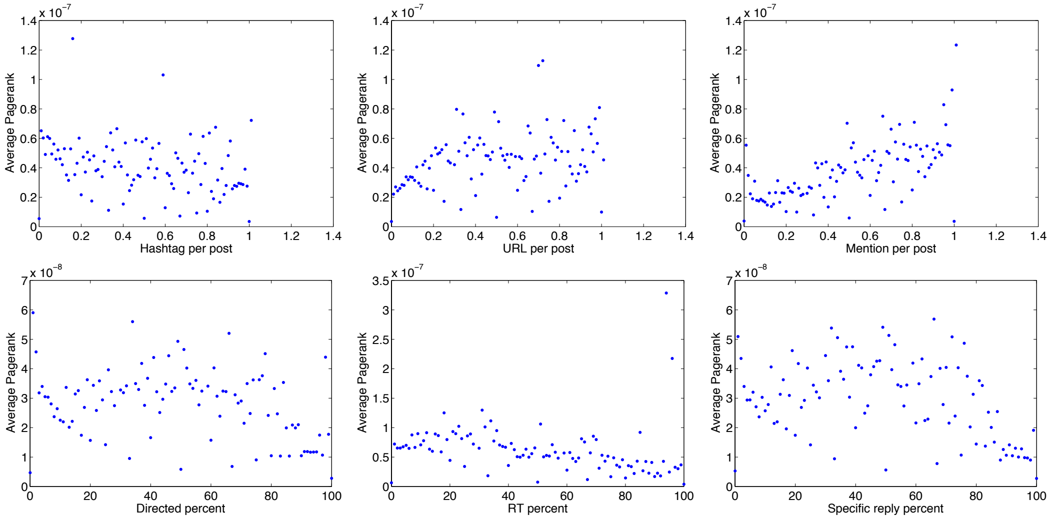


**Figure 14:** Average PageRank for users with varying posting qualities

## 3.2 Parts of the Site

The current functionality of site can be broken down into four major pages. All screenshots of the website show a preliminary design; this may change dramatically by the time the site launches.

### 3.2.1 Home Page

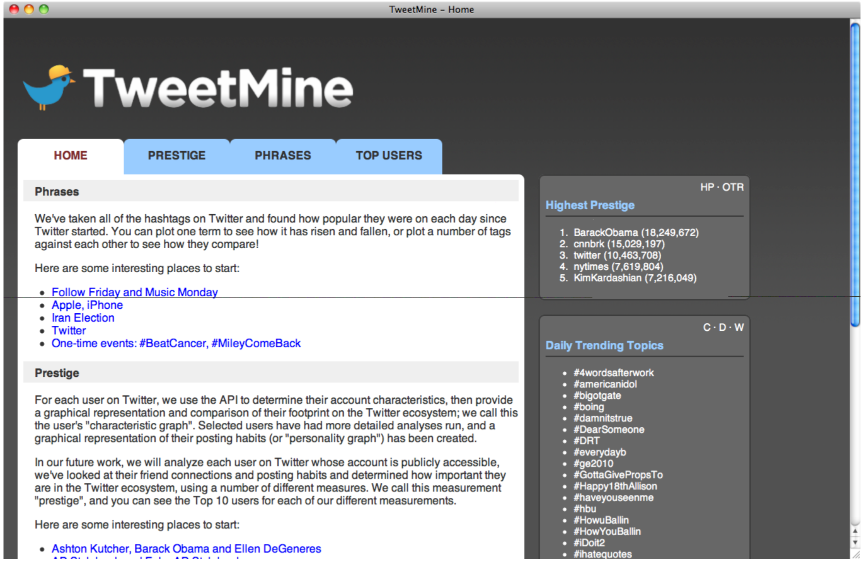
The home page is designed to launch users into the other parts of the site. In particular, it contains links to phrases and prestige pages that we have found to be particularly interesting. It also shows the top 5 most influential users and the current trending topics on Twitter, to engage the user with live circumstances on Twitter.

Figure 15 contains a screenshot of the home page.

### 3.2.2 Prestige

The Prestige section of the site displays information to the visitor about a given set of users. It shows the user's basic characteristics of Followers, Friends and Tweets in numerical form. It also displays this information using a radar plot, to better allow the visitor to understand how significant the differences between the raw numbers are.

The site also displays a "Prestige" value for each user. This value is the user's PageRank [14], scaled so that the total sum of PageRanks is 10 billion, rather than 1. This ensures even users with low PageRank have at least 10 "prestige," and was done in the belief that the average visiter would rather see 10 and $10,000$ rather than $10^{-9}$ and $10^{-6}$ in their comparisons of user prestige.



**Phrases**

We've taken all of the hashtags on Twitter and found how popular they were on each day since Twitter started. You can plot one term to see how it has risen and fallen, or plot a number of tags against each other to see how they compare!

Here are some interesting places to start:

- Follow Friday and Music Monday
- Apple, iPhone
- Iran Election
- Twitter
- One-time events: #BeatCancer, #MileyComeBack

**Prestige**

For each user on Twitter, we use the API to determine their account characteristics, then provide a graphical representation and comparison of their footprint on the Twitter ecosystem; we call this the user's "characteristic graph". Selected users have had more detailed analyses run, and a graphical representation of their posting habits (or "personality graph") has been created.

In our future work, we will analyze each user on Twitter whose account is publicly accessible, we've looked at their friend connections and posting habits and determined how important they are in the Twitter ecosystem, using a number of different measures. We call this measurement "prestige", and you can see the Top 10 users for each of our different measurements.

Here are some interesting places to start:

- Ashton Kutcher, Barack Obama and Ellen DeGeneres

HP · OTR

**Highest Prestige**

1. BarackObama (18,249,672)
2. cnnbrk (15,029,197)
3. twitter (10,463,708)
4. nytimes (7,619,804)
5. KimKardashian (7,216,049)

C · D · W

**Daily Trending Topics**

- #4wordsafterwork
- #americanidol
- #bigotgate
- #boing
- #damnitstrue
- #DearSomeone
- #DRT
- #everydayb
- #ge2010
- #GottaGivePropsTo
- #Happy18thAllison
- #haveyouseenme
- #hbu
- #HowuBallin
- #HowYouBallin
- #iDoit2
- #ihatequotes

**Figure 15:** Screenshot of the Home page

The final part of this page displays “personality” information for users. This is another radar plot, but this time it plots the aggregated post information described in Table 5. This allows the visiter to evaluate what type of poster the user is. A news organization like nytimes or cnnbrk will post many links, but rarely mention another user. A user who uses Twitter to talk to other people probably has many replies and responses, while an account dedicated to finding funny posts of other users will have a large number of retweets.

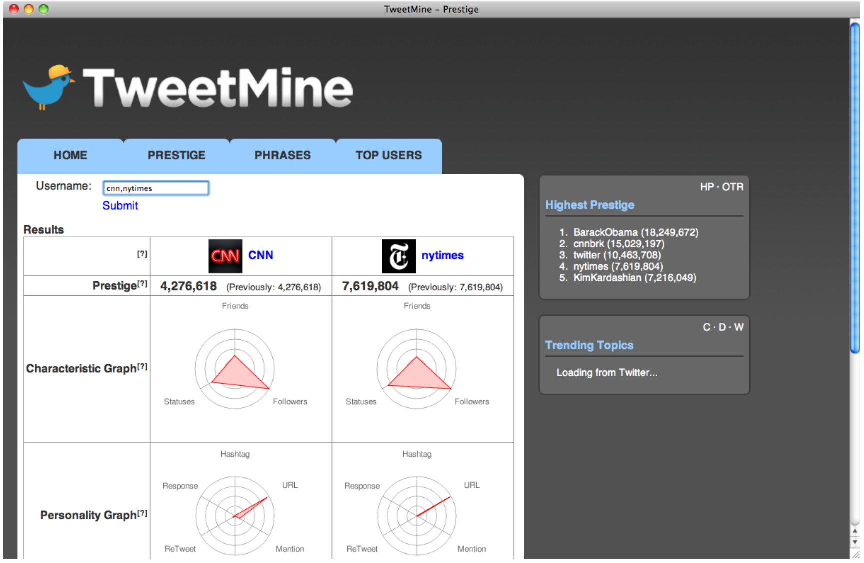
Figure 16 contains a screenshot of the prestige page.

### 3.2.3 Phrases

The Phrases section of the site allows visitors to view the popularity of phrases on Twitter over time. Right now, it is restricted to only hashtags, but this can eventually be extended to general phrases. Visitors search for one or multiple phrases, and an interactive graph appears plotting how many tweets contained those phrases for each day in Twitter’s history. Users can zoom in on parts of the graph to investigate them in more detail.

In future work, we hope to be able to identify the causes of certain spikes in phrase uses on Twitter; for example, the use of the #saints hashtag spiked during the 2010 Super Bowl, and we hope to be able to determine this by analyzing other aspects of the data set. Once this work is complete, the site will be modified to display this event data on the graph timeline.

Figure 17 contains a screenshot of the phrases page.



Username: cnn,nytimes

Results

| [?] | CNN | nytimes |
|---|---|---|
| Prestige[?] | 4,276,618 (Previously: 4,276,618) | 7,619,804 (Previously: 7,619,804) |
| Characteristic Graph[?] | | |
| Personality Graph[?] | | |

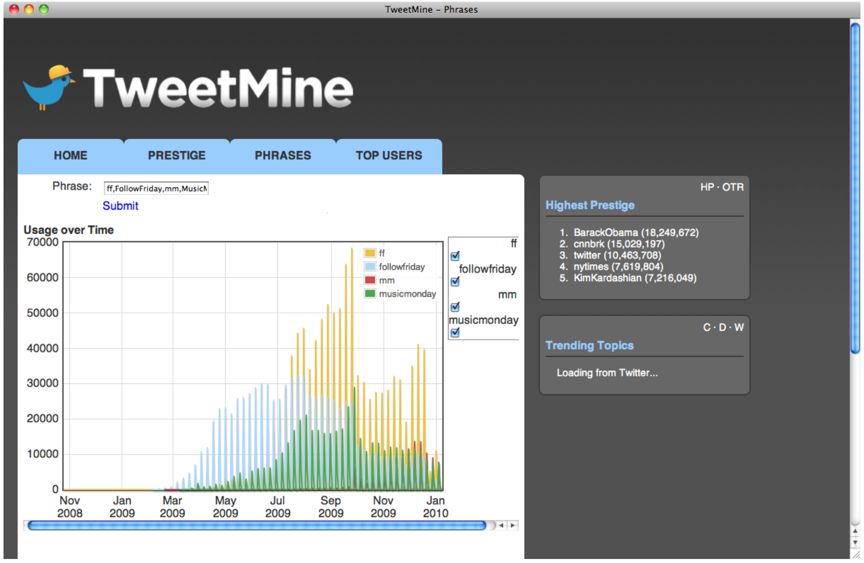
Highest Prestige

1. BarackObama (18,249,672)
2. cnnbrk (15,029,197)
3. twitter (10,463,708)
4. nytimes (7,619,804)
5. KimKardashian (7,216,049)

Trending Topics

Loading from Twitter...

**Figure 16:** Screenshot of the Prestige page



Phrase: ff,FollowFriday,mm,MusicM

Highest Prestige

1. BarackObama (18,249,672)
2. cnnbrk (15,029,197)
3. twitter (10,463,708)
4. nytimes (7,619,804)
5. KimKardashian (7,216,049)

Trending Topics

Loading from Twitter...

**Figure 17:** Screenshot of the Phrases page

### 3.2.4 Top Users

The Top Users section contains two lists of users. The *Highest Prestige* list contains a list of user accounts sorted by their final PageRank; hence, the top user on this list will be the user with the highest total PageRank. The *On The Rise* list contains the users our analysis has suggested are the fastest rising users on Twitter. This will most likely be a function of their PageRank in the previous calculation and their new PageRank; the full details of how this list will be computed is still yet to be finalized.

## 3.3 Impact

The site is still under construction, but should be launched by May 2010. It is our hope that the site will be of interest to the general public, but the work done in the site's creation offer some possibilities for interesting future research.

### 3.3.1 Information Spread

One way in which we hope information about TweetMine will spread is via information spreading on Twitter itself. In particular, if a user searches for just one name in the prestige section of the site, a notice will appear on the page suggesting that they post their results to Twitter. After authenticating their Twitter login info, the site will then use the Twitter API to post a short message containing the prestige results of the user, along with a link back to TweetMine. By examining the referrer information in the server logs of TweetMine, we should be able to learn how many of our visitors went to the site by clicking on a given tweet. Additionally, the user might search for their own username and post their own tweet with information. Based on this data, we now know the following for each user:

- How many followers they have.
- How many of their followers click on links they post.
- How many of those that click the links will then repost the information.

This will hopefully create a large data set about information spreading; combining this data set with the known Twitter social graph will hopefully allow many current models of information spread to be verified.

### 3.3.2 Predicting Popularity

For the Top Users section of the site, an algorithm was developed to predict what accounts are most rapidly gaining in popularity on Twitter; this is the *On The Rise* list. The site logs all of the accounts it predicts for this list, and we can eventually determine how effective various algorithms are at predicting users who did go on to be popular.

Additionally, we are logging the PageRank of all users each time we run the algorithm. After the site has been running long enough, we will be able to plot PageRank over time for various accounts. It is our hope that these plots might be categorizable into different clusters, which will allow us to identify certain types of Twitter users based on their PageRank growth over time.

# 4 Summary and Future Directions

## 4.1 Summary

A large-scale analysis was performed on the Twitter data set, one of the largest social networking data sets available for analysis. It was found that the friend, follower, status and favorite distributions all follow a power-law, though the friend distribution contains anomalies at round numbers. A study of active users on Twitter was performed, and it was discovered that only one-sixth of registered accounts in the network had posted in the last month. Additionally, it was found that larger in and out-degrees increased the activeness likelihood of an account, but that even among accounts with 200 followers, twenty percent were inactive.

The timing of posts was tracked, and plotted over the course of the week. It was found that Monday through Thursday all displayed similar posting tendencies, with a peak usage at around 9:00 PM local time. This discovery dramatically differs from previous studies of web usage over time. The centrality of nodes was determined using the PageRank algorithm, and different properties of the accounts were analyzed to determine their correlation with high centrality. It was found that the contents of a user's posts could not be simply correlated with their centrality, but that users with a certain number of statuses or post frequency had higher centralities on average. In particular, users who post on average once every three hours display the highest centrality.

A website was created to allow the public to view the data collected, as well as the results of the analysis. A page was created to allow the public to view the usage of hashtags on Twitter over time. Another page describes the centrality and aggregated post contents of each user in the data set. Finally, a page was created to display the top users on Twitter, as measured by the centrality analysis. The PageRank data this site provides is currently unavailable elsewhere on the web. The website was crafted to serve as both a useful service for the public and a potential research tool in the future.

## 4.2 Future Work

There are a number of areas of research opened up by the work performed on this data set, many of them addressing the detection of anomalous accounts.

### 4.2.1 Detecting Anomalous Accounts

The analysis of the data set strongly suggests the existence of anomalous accounts in the Twitter social network; Twitter itself acknowledges their presence, and is working to remove them [5]. The analysis performed as part of this thesis can be used to help detect which accounts are anomalous, and to determine if a given automated account is a spammer.

To formalize our discussion, we need to establish what exactly an anomalous account and a spammer are. We will define an anomalous account to be “any account whose content is not entirely generated by a human.” For example, the *New York Times* Twitter account (nytimes) is anomalous; a post is automatically generated for various articles from that newspaper.

We can define a spammer using the same definition used by Twitter in [6]. Examples of spam behavior defined there are posting harmful links, posting links with unrelated tweets, or using trending topics to grab attention. While the *New York Times* Twitter account is anomalous, it would not be classified as spam under this definition. On the other hand, we can probably make the assumption that all spam accounts will be anomalous; hand-crafting spam messages, if feasible, is not a large enough issue compared to computer-generated spam.

**Ground Truth** A key difficulty in this analysis is establishing a ground truth for whether an account is anomalous or not, and whether an anomalous account is a spammer. There could be two main techniques for establishing this.

**Twitter suspension** Our data set was gathered in November of 2009. In rescrapes performed since that time, some accounts could not be rescraped, as they had been suspended by Twitter. We will thus assume that any account suspended by Twitter was anomalous, and was a spammer. The converse does not hold, however; accounts not suspended by Twitter are certainly not guaranteed to be non-anomalous accounts.

**Human Computation** We hence would need another technique to detect those anomalous accounts remaining on Twitter that have not been shut down. One could set up a system using Amazon Mechanical Turk [2] to use human computation to establish whether an account is anomalous. To do so, the Mechanical Turk user could be shown a sampling of the user’s tweets, then asked one of two questions:

- Are this account’s posts being generated by a computer?
- Is this account a spammer?

Based on the responses to these questions, we can determine ground truth for whether a user is anomalous or not, and whether they are a spammer or not.

**Detecting Anomalous Accounts with PageRank** The PageRank system proved extremely effective for Google in determining whether webpages were important or not. It is our hypothesis that running PageRank on the Twitter social graph gives high values to legitimate users, and lower values to anomalous accounts. In addition, the difference should be even more dramatic when the PageRank is run with the random surfer visiting only known valid accounts (those marked as "verified accounts" by Twitter). The difference should also be emphasized when edges are created only when @-mentioning has occurred, and not just when following occurs.

Additionally, the work in Subsection 2.6 found that accounts with certain posting properties have a higher average PageRank. An account which meets many of these properties, but has a dramatically lower PageRank than expected might be a good candidate for an anomalous account.

**Detecting Anomalous Accounts with Machine Learning** In this paper, a number of characteristics of users were discussed: the six aggregate post measures from Table 5; friend, follower and status counts; post frequency and account age are simple ways to characterize a given user. With these areas, we can take each user and create a vector representing that user. We could then perform unsupervised learning to try and cluster the anomalous accounts. Alternately, we could establish ground truth on a training set with the techniques described above, then use supervised learning to train a classifier on detecting spammers.

## 4.3 Acknowledgements

I'd like to begin by thanking Luis von Ahn, my thesis advisor. Both my interest in theoretical computer science and in social network theory can be directly traced back to courses I took with him. His advice on this project, research, academia and everything in between is truly appreciated, and my experience at Carnegie Mellon would have been greatly diminished without his guidance.

The importance of Brendan Meeder to this thesis cannot be understated. The idea of scraping Twitter for this data set was his, and he has helped with virtually every aspect of this thesis. The first section of this paper was taken from a conference paper I co-authored with him, and his suggestions on how to proceed with this project led to many of the interesting problems and conclusions of the thesis.

Some of the ideas in this report were inspired by discussions with Jon Kleinberg of Cornell and Jure Leskovec of Stanford. Their helpful suggestions and advice are greatly appreciated. Twitter has chosen to maintain an open and easily accessible service, enabling this research to take place. They have also also been helpful in working out issues with the crawler and their API. Gathering the data set for this thesis required an large amount of computational resources, which were provided by Adrien Treuille and the CMU Graphics Lab.

Finally, thanks to my friends and family, whose support has been invaluable not just in the completion of this thesis, but throughout my entire life.

# References

[1] About twitter. http://twitter.com/about. Accessed 20 Oct 2009.

[2] Amazon mechanical turk – welcome. https://www.mturk.com/mturk/welcome. Accessed 18 Apr 2010.

[3] Twitter api wiki / twitter rest api method: statuses retweet. http://apiwiki.twitter.com/Twitter-REST-API-Method:-statuses-retweet. Accessed 20 Oct 2009.

[4] Twitter api wiki / twitter rest api method: statuses retweets. http://apiwiki.twitter.com/Twitter-REST-API-Method:-statuses-retweets. Accessed 20 Oct 2009.

[5] Twitter blog : State of twitter spam. http://blog.twitter.com/2010/03/state-of-twitter-spam.htm
Accessed 23 Apr 2010.

[6] Twitter support : How to report spam on twitter. http://help.twitter.com/entries/64986. Accessed 18 Apr 2010.

[7] Y. Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. Feb 2009.

[9] L. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[10] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[11] U. Kang, C. E. Tsourakakis, and C. Faloutsos. Pegasus: A peta-scale graph mining system- implementation and observations.

[12] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.

[13] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA, 2007. ACM.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[15] R. Wauters. Url shortening wars: Twitter ditches tinyurl for bit.ly. `http://www.techcrunch.com/2009/05/06/url-shortening-wars-twitter-ditches-tinyurl-for-b` May 2009. Accessed 20 Oct 2009.