

A Few Chirps About Twitter

Balachander Krishnamurthy
AT&T Labs – Research
bala@research.att.com

Phillipa Gill
University of Calgary
pssessini@ucalgary.ca

Martin Arlitt
HP Labs/University of Calgary
martin.arlitt@hp.com *

ABSTRACT

Web 2.0 has brought about several new applications that have enabled arbitrary subsets of users to communicate with each other on a social basis. Such communication increasingly happens not just on Facebook and MySpace but on several smaller network applications such as Twitter and Dodgeball. We present a detailed characterization of Twitter, an application that allows users to send short messages. We gathered three datasets (covering nearly 100,000 users) including constrained crawls of the Twitter network using two different methodologies, and a sampled collection from the publicly available timeline. We identify distinct classes of Twitter users and their behaviors, geographic growth patterns and current size of the network, and compare crawl results obtained under rate limiting constraints.

Categories and Subject Descriptors

C.4 [Performance of Systems]: [Measurement techniques, Modeling techniques]

General Terms

Measurement, Performance

Keywords

Online Social Networks, Measurement

1. INTRODUCTION

Online social networks (OSNs) have emerged recently as the most popular application since the Web began in the early 1990s. Coincident with the growth of Web 2.0 applications (such as mashups, user generated content) and users being treated as first class objects, numerous social networks along with thousands of helper applications have arisen. Well known ones include Facebook, MySpace, Friendster, Bebo, hi5, and Xanga, each with over forty million [13] registered users. Many applications have been created to use the distribution platform provided by OSNs. For example, popular

* Author ordering is reverse alphabetical.

games like Scrabulous, allow many thousands of users on Facebook to play the game with their social network friends. A few smaller networks with superficial similarities to the larger OSNs have started recently. Some of these began as simple helper applications that work well with the larger OSNs, but then become popular in their own right.

A key distinguishing factor of these smaller networks is that they provide a new means of communication. In the case of Twitter [21] it is Short Message Service (SMS [18]), a store and forward best effort delivery system for text messages. In the case of *qik*, it is streaming video from cell phones. Jaiku [10], another small OSN, allows people to share their “activity stream”, while Dodgeball [6] lets users update their status along with fine-grained geographical information, allowing the system to locate friends nearby. GyP-Sii [8], a Dutch OSN is aimed at the mobile market exclusively, combining geo-location of users with image uploading and works on various cell phones including Apple’s iPhone. Close to Twitter, a mobile OSN that encourages constant updates is Bliin [3]. Other examples of exclusively mobile social networks include Itsmys and MyGamma.

A distinguishing factor of such smaller networks and applications is their ability to deliver the data to interested users over multiple delivery channels. For example, Twitter messages can be received by users as a text message on their cell phone, through a Facebook application that users have added to their Facebook account to see the messages when they log in, via email, as an RSS feed, or as an Instant Message (with a choice of Jabber, GoogleTalk etc.). Figure 1 shows the various input and output vectors to send and receive Twitter status update messages (“tweets”). Twitter is an example of a *micro-content* OSN, as opposed to say, YouTube, where individual videos uploaded are much larger. Individual tweets are limited to 140 characters in Twitter.

Twitter began in October 2006 and is written using Ruby on Rails [16]. Our study finds that users from a dozen countries are heavily represented in the user population but significantly less than the U.S. Recently, Twitter has made interesting inroads into novel domains, such as help during a large-scale fire emergency [4], updates during riots in Kenya [1], and live traffic updates to track commuting delays [20].

Our goal is to characterize a novel communication network in depth, its user base and geographical spread, and compare results of different crawling techniques in the presence of constraints from a generic measurement point of view. Section 2 presents the details of our various crawls of the Twitter network. Section 3 presents a detailed characterization of the Twitter network. We explore related work and conclude with ongoing work in Section 4.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WOSN’08, August 18, 2008, Seattle, Washington, USA.
Copyright 2008 ACM 978-1-60558-182-8/08/08 ...\$5.00.

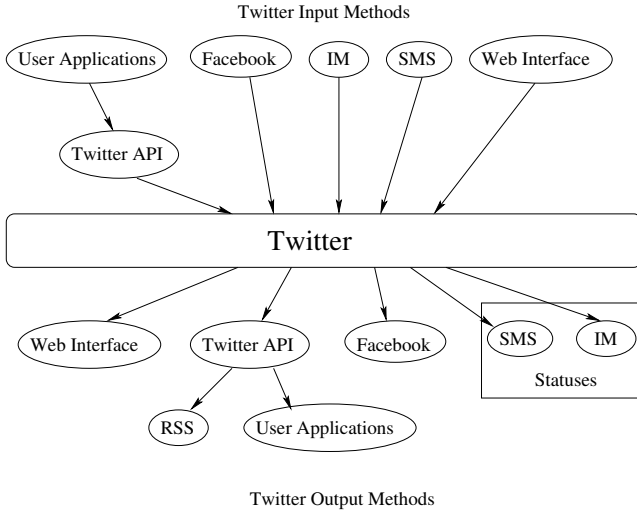


Figure 1: Twitter input and output methods

2. STUDY METHODOLOGY

We used two main data collection methods, both relying on the API functions provided by *twitter* [2]. We gathered detailed information on the users and the list of users each of them were *following*. The constraint on the number of queries that we could issue in a day was the key limiting artifact in the reach of our crawl. A Twitter user interested in the statuses of another user signs up to be a “follower”. Relationships in Twitter are directed but there are no methods available for gathering the set of reverse links (i.e., the set of users following a user). We also use the “public timeline” API method that returns a list of the 20 most recent statuses posted to *twitter.com* by users with custom profile pictures and unrestricted privacy settings.

The first dataset (“crawl”) gathered by a constrained crawl of the Twitter network, was seeded by collecting the public timeline at four distinct times of day (2:00, 8:00, 14:00, and 20:00 Mountain Time) and extracting the users that posted the statuses in these timelines. Each step in the crawl involved collecting details of the current user as well as a *partial* list of users being followed by the current user. During this process the median number of users followed by the previously crawled users, m , was tabulated [14]. To further the crawl the first m users followed by the current user would be added to the set of users to crawl. If the current user followed fewer than m users, all users are added to the set of users to crawl. It should be noted that while the users that posted statuses are clearly currently active, the list of users obtained in successive steps may not have been active. This first dataset is likely to include a certain fraction of passive users. The duration of data gathering was three weeks from January 22 to February 12, 2008 and information about 67,527 users was obtained.

The second dataset (“timeline”) was gathered via the public timeline command to sample currently active *twitter* users. Twitter continually posts a series of twenty most recent status updates. Samples were made by retrieving the public timeline and extracting the set of users associated with the statuses in the timeline. Details of these users were then collected. Once details of the users from the previous timeline were gathered the public timeline was queried again to find the next set of users. This process was repeated for a period of three weeks (Jan. 21 to Feb. 12, 2008) resulting in samples from various times of day and days of the week. Information about 35,978 users were gathered in this dataset.

Finally, to examine potential bias in our constrained crawl, an additional dataset of 31,579 users was gathered between February 21–25, 2008, via the Metropolized random walk with backtracking, used for unbiased sampling in P2P networks [19]. Note that this crawl required fewer requests as we considered only one child of each node and the rate limiting was slightly relaxed. Our analysis presents results on all the datasets with comparisons as warranted.

3. CHARACTERIZATION RESULTS

We present analysis in four parts: characterization of Twitter users, status updates, validation of the crawl methodology, and some miscellaneous insights.

3.1 Characterization of Twitter users

With nearly 100,000 users in the three datasets combined, we believe that we can extract broad attributes of Twitter users. We begin by examining the number of users each user follows and the number of users they are followed by, to get an idea of the nature of connections between users in micro-content social networks.

The relationship between the number of followers and following is explored in Figures 2–4. Figure 2 shows a scatter plot of the follower/following spread in the crawl dataset. Three broad groups of users can be seen in this figure. The first group appear as vertical lines along the left side of Figure 2. These users have a much larger number of followers than they themselves are following. This behavior characterizes *broadcasters* of tweets. Many of the users here are online radio stations, who utilize Twitter to broadcast the current song they are playing. Others include the New York Times, BBC, and other media outlets generating headlines.

A second group of users labeled *acquaintances*, tend to exhibit reciprocity in their relationships, typical in online social networks [15]. Users in this group appear in the large cluster that falls (roughly) along the line $y=x$ in Figure 2.

A third unique group of users is a small cluster around the line $x=7000$ in Figure 2. A common characteristic of these users is that they are following a much larger number of people than they have followers. Such behavior is typical of *miscreants* (e.g., spammers or stalkers) or *evangelists*, who contact everyone they can, and hope that some will follow them [9]. We continue to work on a better characterization of this evolving group. For example, one month after the crawl data was collected, one of the users in this group has increased his following count from 7,462 to 31,061¹. The top datapoint on $x=7000$ is John Scoble, a technical blogger who follows roughly 70% of the people who follow him.

The vertical lines corresponding to $x=1, 2, \dots, 10$ in Figure 2 happen to be broadcasters as well who are following the primary broadcaster at $x=0$. For example, a top broadcaster somafm illstreet (140,183 updates) has 213 followers, and is following 11—all of whom are sister radio stations.

Figure 3 shows the ratio of followers and following for all three datasets. This figure indicates that the groups identified in Figure 2 appear in all three datasets. The bulk of the users exhibit roughly symmetric behavior. The head and tail of the distribution reflect the evangelists/miscreants and broadcasters, respectively.

Next we examine the relationship between the number of status updates (“tweets”) and the following/follower relationship. Figure 4 contains three sets of data points. The “all” data points plot the following/follower relationship for all users in the crawl data (same as Figure 2). The “90%” and the “99%” data points plot the following/follower relationship for the top 10% (90th percentile - 964

¹Over this same period, his number of followers has decreased from 3,333 to 3,260.

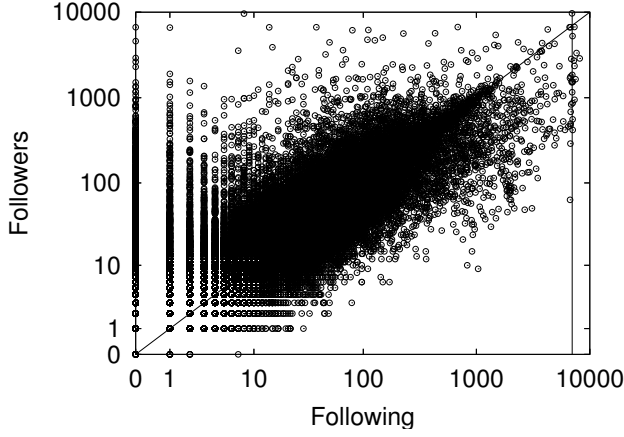


Figure 2: Scatterplot of crawl users' following and follower count

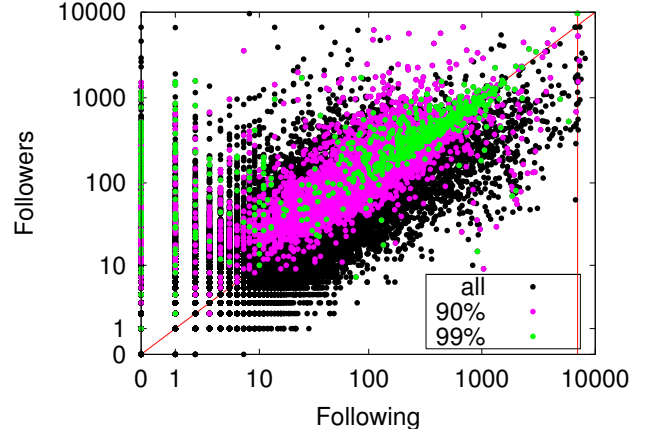


Figure 4: Scatterplot of crawl users' following and follower count, by status update level

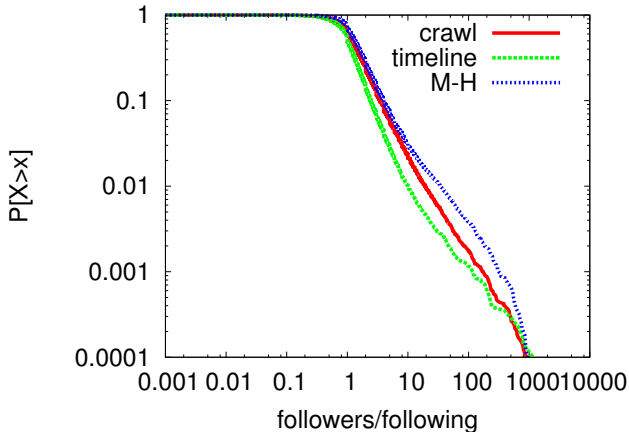


Figure 3: CCDF of users' following and follower count

or more tweets during the user's lifetime) and the top 1% (99th percentile - 1,727 or more tweets) of tweeters.

Figure 4 shows that many of the users in the first group do tweet a lot, confirming that they are broadcasters. In the acquaintances group, an interesting characteristic is that the following/followed relationships move closer and closer to complete reciprocity as the number of tweets increases; looking at the 99% data points, most of them fall reasonably close to the diagonal. Lastly, we find that most of the members in the third group are not among the top tweeters.

Twitter users can include their URL information; both URL and the UTC offset are present in nearly two thirds of users in crawl and timeline datasets. Comparing the domain information in the URLs with the UTC offset allows us to see popularity of Twitter in different countries. Users with URL in the .com domain are largely likely to be in North America but the UTC showed some of them to be in Europe as well. Beyond this, the rest of the UTC data lined up with the domain information. After USA, the top-10 countries are Japan, Germany, U.K., Brazil, Holland, France, Spain, Belgium, Canada, and Italy. These eleven countries account for around 50% of users in our datasets.

Table 1: Twitter Sources

Crawl		Source	Timeline	
%	Statuses		%	Statuses
61.7	40,163	Web	57.0	20,510
7.5	4,901	txt (mobile)	7.4	2,667
7.2	4,674	IM	7.5	2,714
1.2	792	Facebook	0.7	261
22.4	14,566	Custom Applications	27.3	9,821

3.2 Characterization of status updates

We also examine the source interface used for posting Twitter messages in Table 1. The distribution of sources are nearly identical in crawl and timeline datasets with the top dozen sources accounting for over 95% of all tweets. Nearly 60% come from "Web" which includes the *twitter.com* Web site and unregistered applications that use the API. Mobile devices and Instant Messages have visible presence. A fifth of all status updates come from the various custom applications that have been written using the Twitter API. Twitter traffic increased significantly when the API was opened up [17]. The custom applications are for different OSes (e.g., *twitterrific* for Macintosh, *twitterwindows* for Windows in Japanese), browsers (*twitterfox* for Firefox), RSS feeds/blogs (*twitterfeed*, *netvibes*, and *twitter tools*), desktop clients (*twihirl*, *snitter*), OSNs (*Facebook*), and mobile clients (*movatwitter*), and Instant Message tools.

Figure 5 shows the time of day when status updates are posted (adjusted to local time of the updaters). There is no significant difference between the crawl and timeline datasets. The workload shows a rise during later morning hours, relatively steady use throughout the day, and drop off during the late night hours. There was no significant information in the patterns within days of the week (not shown). Also not shown, there is virtually no difference between the length of tweets in the crawl and timeline datasets.

3.3 Comparing the datasets

Our methodology to gather Twitter data had a key constraint: we were limited by the Twitter user agreement in the number of requests we could issue each day. Yet, we were able to gather data about over 67,000 users via our crawl. At the same time we were able to fetch public timeline data made available by Twitter.

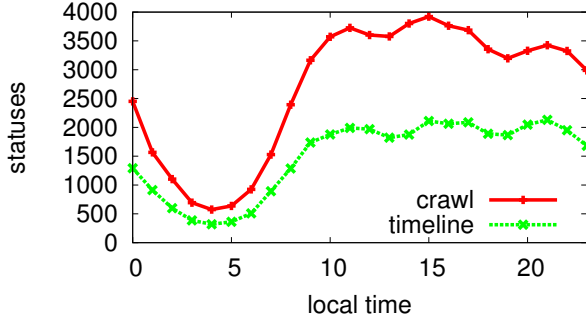


Figure 5: Time of day status update of crawl and timeline datasets

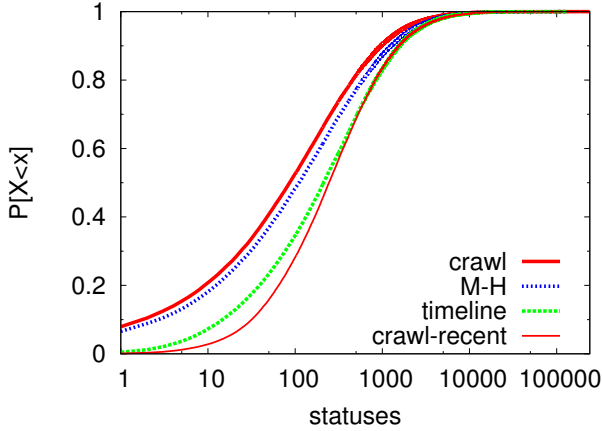


Figure 6: CDF of count of statues comparing crawl and M-H with timeline

Drawing inferences about the global Twitter graph depends on the representativeness of the portion of the graph we have captured. The status updates in the timeline dataset are presumably a random snapshot of currently active users. As mentioned in Section 2 the crawl dataset could include users who have not been active recently. The representativeness of the crawl requires correction for bias towards high degree nodes; adding backtracking to the random walk [19] is one way. We implemented the Metropolized random walk variant in the data collection and gathered the M-H dataset of over 31,000 users. The Metropolized random walk ignores the semantics of any particular graph. The connection model of the Twitter graph differing from a graph of users who exchange data in P2P networks should not impact us.

In the rest of this section we compare various characteristics of the three datasets and see if differences can be explained based on our additional knowledge of the semantics of the Twitter application and its user population.

Figure 6 shows that the Metropolized random walk algorithm yields a portion of the Twitter graph that has nodes with very similar status count as the crawl dataset. Both have fewer statuses as compared to the active nodes represented in the timeline dataset. To confirm this, we examined the portion of users in the crawl data who tweeted *during* our data gathering—they also had a higher count of statuses.

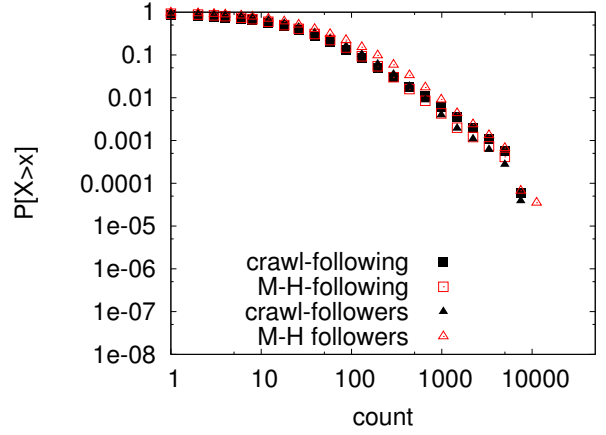


Figure 7: CCDF of followers and following count for the three datasets

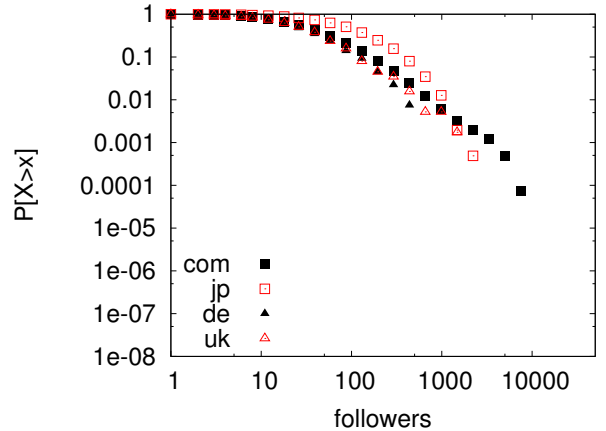


Figure 8: CCDF of followers count for the top 4 domains

Figure 7 shows the overall similarity of results between crawl and M-H datasets in the CCDF of the count of followers and following. M-H has slightly more followers.

Figures 8 and 9 show the CCDF of followers and following for the data restricted to users in the top four domains **.com**, **.jp**, **.de**, and **.uk** in the crawl dataset. Although this is somewhat similar to Figure 2(b) in [15] (LiveJournal indegree and outdegree graph), we prefer to make comparisons within our dataset as we understand the Twitter milieu better and we want to stray from the conventional power law result. A higher friends and followers count can be seen in the **.jp** domain, perhaps reflective of the more connected nature and popularity of such technologies in Japan.

Our datasets include several additional fields on each user including *location* and *utc_offset*. Both of these present clues to the geographical presence of the user. Comparing the crawl and timeline dataset with respect to these fields will also show representativeness of the crawl dataset. We examined the UTC offset attribute of each user. Figure 10 shows the percentage of users in each UTC offset in the crawl and timeline datasets. As can be seen there are many more users in the Japan timezone not captured in the crawl dataset as compared to the timeline dataset. There is also a cultural separation to a certain, expected, degree. Users with UTC of GMT+9 indicates a large group of users in the **.jp** domain. They use

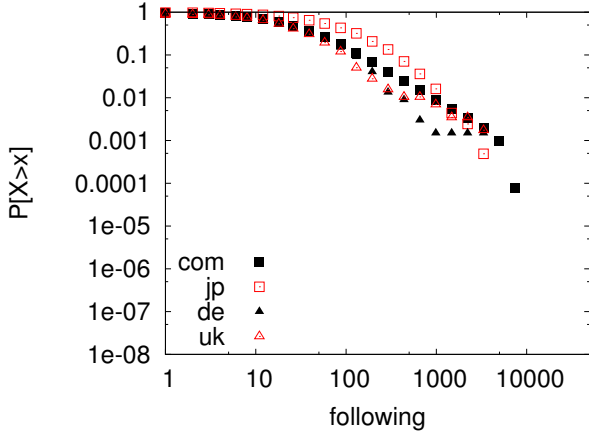


Figure 9: CCDF of following count for the top 4 domains

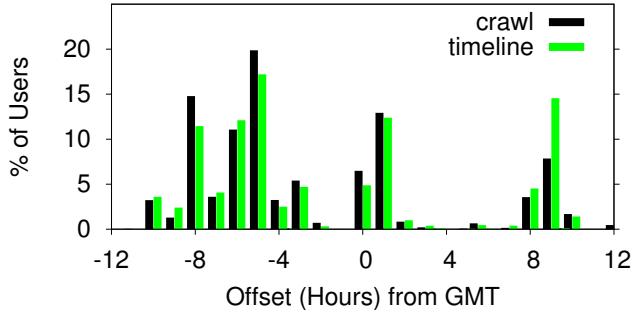


Figure 10: Histogram of UTC offset hours comparing crawl and timeline

Japanese to communicate with each other, leaving out most of the English language tweeters. Similarly there are (smaller) clusters of German, Italian, etc. users who tweet to each other.

3.4 Other properties of Twitter

We examined if highly popular users (those who have many followers) update their status more often than those who (likely passively) follow more users. This was true in both the crawl and timeline (not shown) datasets. Figure 11 shows that crawl dataset users who have more than 250 followers send many more status updates than those who follow more than 250 users. The 250 cutoff value was chosen as it was just above the 95th percentile in both datasets.

We tried to estimate the approximate number of Twitter users based on the integer identifiers assigned to them. Figure 12 shows, for the crawl and timeline datasets, binned per thousand. Twitter appears to assign all numbers in two small ranges, else they have only been assigning 1/10th of the unique integers. From the crawl and timeline datasets we can see that they have used all numbers between 0 and 13,743 and then switched to 3 mod 10. They then switched back to sequential assignment until around 754,363, then to 1 mod 10 at around 825,000, and to 2 mod 10 at around 5,283,000. The largest userID in crawl is 12,978,372 and timeline is 13,389,452. This allows us to estimate the total number of users around 1.4 Million at the time of data collection. We verified this by constructing figures (not shown) for users in Japan, Europe, U.K., U.S.A. (East and West coast timezones) using the UTC off-

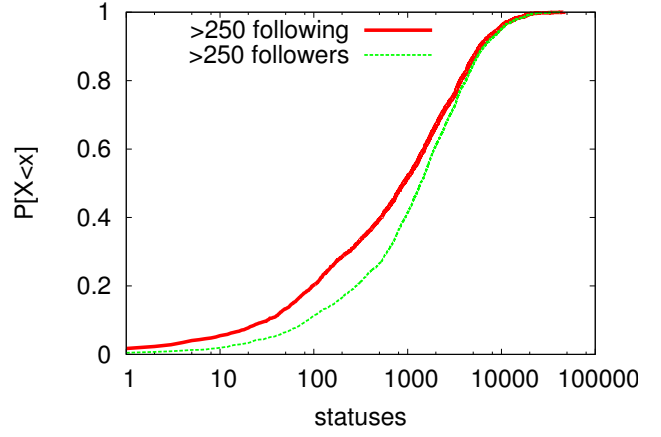


Figure 11: CDF of heavy users and followers in crawl

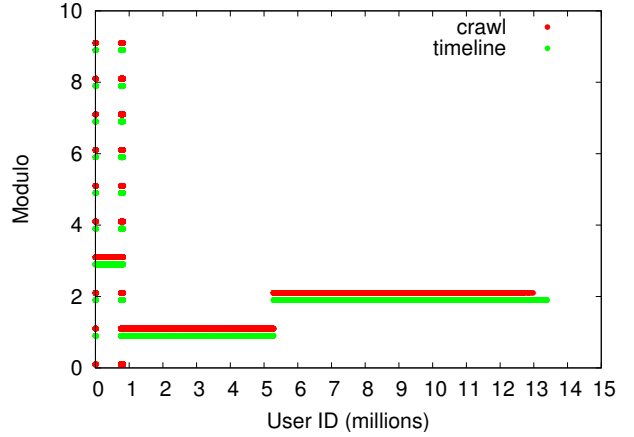


Figure 12: Estimate of Twitter users based on userids

set information—all the figures were identical with respect to the distribution of userIDs.

Finally, we look at geographical growth of Twitter users. Figure 13 shows for the union of users in crawl and timeline dataset that had a UTC offset (98.5%), the growth of users in each distinct geographic region over time. Although Twitter was adopted in Japan later, it has grown quickly to become the third largest region. Asia-Pacific region includes everyone not covered by the top four regions.

4. RELATED WORK AND SUMMARY

An earlier examination of Twitter usage [11] has drawn similar inferences in followers and following counts, different classes of users, and symmetry of relationships. However, our study is broader and improves on their work in several ways. First, [11] uses data from a single source (the public timeline); we use three different data collection techniques and examine their strengths and weaknesses. Second, they assumed sequential growth in userIDs; we demonstrated that this is not the case. Third, we factor in tweet count to show heavy tweeters tend to have a more reciprocal relationship. Fourth, we use both the top-level domain and UTC offset to identify location of a much larger fraction of users; we also ex-

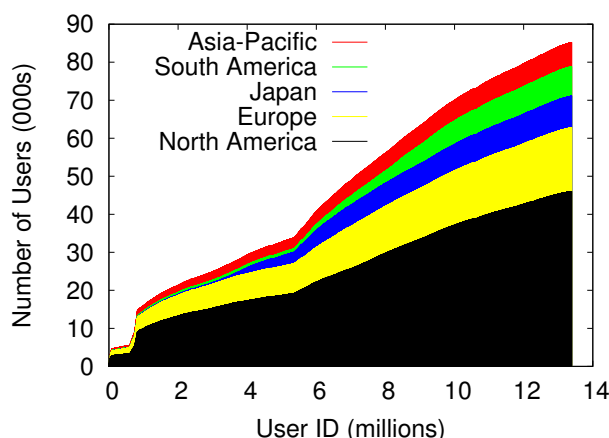


Figure 13: Geographical presence of Twitter based on UTC data

amine the growth of users by geography. In addition, we examine number of tweets/user, time of day use, sources of tweets, and distribution of userIDs.

Work on regular OSNs range from characterization, analysis, to comparing different OSNs. Given both the smaller size of Twitter and the minimal overlap with the features available on larger OSNs, direct comparisons are risky. However, we expect the smaller networks to grow in different directions and the large established base of cell phone users are likely to participate in OSNs using mobile devices.

Several popular OSNs have been studied recently. A study [12] of Flickr and Yahoo! 360 networks examined path properties (such as diameter), density (ratio of undirected edges to nodes) change over time, and presence of a giant component. One point in common appears to be that a few people choose to engage more deeply in interactions—this is true among the human users of Twitter. Flickr, Orkut, LiveJournal, and YouTube were studied on a reasonably large scale [15], with inferences relating to the small-world nature. They showed the presence of symmetry in link structure in terms of in- and outdegrees which we see for a reasonable portion of Twitter users. However, in Twitter there are some high degree nodes due to the presence of broadcasters. There are a few high follower users, namely Web celebrities. Twitter does not appear to have any visible limits on the number of friends/followers unlike LiveJournal or Orkut. YouTube was studied with an emphasis on characterizing user generated content [5, 7]. Properties of the content, popularity distributions and strategies for handling the resource demands of an OSN that centers around large content were considered. Unlike YouTube, Twitter centers on very small content and presents different challenges to systems design.

In conclusion, we examined geographical distribution, the user base of a new, popular, micro-content network. We compared the results of our constrained crawl against other datasets to show similarities in results. We are examining the shift in Internet traffic towards program or machine generated data and consumption by processes or filters on behalf of human users. The explosion of automatic generators is likely to lead to further split traffic streams.

5. REFERENCES

- [1] A-twitter - a hip version of cablese jazzes up campaign coverage. http://www.economist.com/world/na/displaystory.cfm?story_id=10608764.
- [2] About twitter's api. <http://twitter.com/help/api>.
- [3] bliin! YourLive! Always there. <http://www.bliin.com>.
- [4] California fire followers set twitter ablaze. <http://blog.wired.com/monkeybites/2007/10/california-fire.html>.
- [5] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I tube, you tube, everybody tubes. In *IMC*, 2007.
- [6] Mobile social software. <http://www.dodgeball.com>.
- [7] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: A view from the edge. In *IMC*, 2007.
- [8] Gypsii webtop. <http://www.gypsii.com>.
- [9] Is twitter about to have a big spam problem? <http://mashable.com/2008/03/24/twitter-spam/>.
- [10] Jaiku: Your conversation. <http://www.jaiku.com>.
- [11] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *KDD*, 2007.
- [12] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD*, 2006. Research Track Poster, KDD.
- [13] List of social networking sites. http://en.wikipedia.org/wiki/List_of_social_networking_websites.
- [14] G. S. Manku, S. Rajagopalan, and B. G. Lindsay. Approximate medians and other quantiles in one pass and with limited memory. *SIGMOD*, 27(2):426–435, June 1998.
- [15] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC*, 2007.
- [16] Web development that doesn't hurt. <http://www.rubyonrails.org/>.
- [17] APIs increase twitter's traffic 20x. http://blogs.guardian.co.uk/digitalcontent/2007/11/silicon_valley_comes_to_oxford_1.html.
- [18] Short message service. http://en.wikipedia.org/wiki/Short_Message_Service.
- [19] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peertopeer networks. In *IMC*, 2006.
- [20] Track commuting delays via twitter with commuter feed. <http://lifesacker.com/355453/track-commuting-delays-via-twitter-with-commuter-feed>.
- [21] Twitter: What are you doing? <http://www.twitter.com>.