Twitter Locality:

Analyzing Communication Patterns in Social Networking

Joseph Noor
Department of Computer Science
University of California, Los Angeles
jnoor@cs.ucla.edu

ABSTRACT

In recent times, the widespread use of the Internet has led to the rise of social networking as a means for inexpensive and rapid message passing. This has opened up communication pathways across the globe, overcoming the physical distance between any two users. In this paper, we examine the extent to which this barrier has been broken. By observing a random sample of tweets over the course of one week, we construct a graph of communication links spanning across the globe. We then identify representative and statistically significant trends within the data in an attempt to characterize the different users and locations with respect to physical distance. A lognormal model is shown to reasonably approximate global communication. Key characteristics such as popularity, country of origin, and time of communication are all shown to influence the range of communication. From the results and graphs shown, it is clear that the omnipresence of the internet has allowed for a worldwide communication bandwidth that is much greater than has ever been seen before, and it is useful to identify which countries are actually taking advantage of this.

1. INTRODUCTION

The emergence of computers and the Internet has marked the beginning of what is referred to as the "Information Age," in which innovation in Information and Communications Technology has allowed for significant technological advancements [1]. One of the most notable changes to come about from the Information Age is enabling mankind to connect and send messages across the entire world. Previously, the only means to communicate across long distances was through telecommunications. However, in many parts of the world, telecommunications remain expensive, remote, and unreliable. The Internet, on the other hand, is unique in that it allows for fast and cheap communication anywhere, given that the sender and receiver have the means to connect to the web. In essence, the Internet provides a global medium by which anyone connected is able to send and receive messages [2].

Recent studies have shown that around 2.7 billion people, 40% of the world population, have access to Internet. The most penetrative means by which individuals communicate online is through social networking sites. Websites such as Facebook [6], Twitter [9], and Google+ [13] have become massively popular over the last decade because they provide this over-the-web transmission to their users for little to no cost. In fact, social

networking sites have gained such incredible popularity that social scientists have dubbed the current generation the "Social Media Generation," due to the fact that users spend a very large percentage of their time connected to these social networking sites [3].

The implications of the rise of social networking are numerous. Communications studies, sociology, marketing, macroeconomics, and many other social sciences have all undergone major changes due to this phenomenon. One intuitive question that arises from observing the prevalence of the Internet is, "how are users actually communicating?" In this paper, by analyzing current communication patterns, the answer to this question will be revealed. Furthermore, we aim to categorize and understand the distribution of communication over the Internet. In particular, the focus of this paper is on discovering patterns of communication with respect to physical distance, for this is the fundamental limitation that the Internet has been able to supersede.

There are many challenges that arise when attempting to analyze and model social networking communication patterns. First and foremost is the overwhelmingly large size of the web, its users, and its traffic. As an example, Facebook has over one billion active users [7] with more than one billion messages sent every day [8]. Of course, attempting to collect all of this data in order to truly capture the complete distribution is near impossible. Furthermore, due to the proprietary nature of Facebook and other social networking websites, access to this type of information is heavily restricted, such that the average user would have no way of obtaining this data. Since we are looking to understand how communication ranges over different distances, it is important to have specific location information for each individual user. Finally, managing such a potentially large amount of communication data involves using modern techniques, such as database management systems, to efficiently organize and process said data.

This paper is organized into five main sections. In Section 2, the challenges above are solved and/or mitigated. The metric used to rank distance between users is given. Also, the setup for the data collection and analysis is presented, such that any reader could easily replicate the data collection. In Section 3, the means by which data was collected is described. The raw data is also shown, along with any initial observations regarding the raw dataset. In Section 4, patterns and trends in the data are identified and compared, allowing for the characterization of different types of social networking communication. Also, the distribution of communication is shown to follow a lognormal model. Finally, in the remaining sections, concluding observations are made and our initial questions are answered.

There currently exists no prior work regarding the spatial locality of social network communication. Previous work in understanding social communication trends led to a Social Network Analysis of the web as a graph with a "distance" metric equating to frequency of communication [4]. However, never before has physical global distance been used as a metric to evaluate and categorize communication trends. Intuitively, one may propose the hypothesis that, in fact, most communication is heavily localized. Even if this hypothesis is true, it is not the interesting question. Instead, it is important to understand how different users are communicating, and why.

2. SETUP

In order to make any sort of conclusions regarding the type of communication on the web, the initial and fundamental step is to actually possess the communication data. After failed attempts at gaining access to Facebook's message logs, it was through Twitter's API that these types of messages were collected. Storing all of the communication data present throughout the history of Twitter requires access to a datacenter that only companies such as Twitter and Facebook possess. However, by obtaining a large enough random sample of all communication, it is possible to extrapolate the results from the sample to estimate total communication [5]. This is the approach taken in this paper, using Twitter's Streaming API [10].

2.1 Mini-Twitter and Random Sampling

A snapshot of Twitter is taken over one week, dubbed "Mini-Twitter." By finding characteristics in Mini-Twitter, we can extrapolate the conclusions to all of Twitter communication, and all social networking communication, within a reasonable margin of error. Even though Twitter's API is heavily restrictive, it is possible for the average Internet user to obtain similar data by using Twitter's Streaming API. As Twitter receives tweets, the Streaming API forwards these tweets to applications connected to the API.

Twitter offers two types of connections to their Streaming API, dubbed "firehose" and "sample." The firehose connection forwards all incoming tweets, but requires special permission to access. After a failed attempt to gain access to the firehose, the remaining option was to use the sample stream. According to the Twitter API, connecting to the sample stream will forward a random sample of all tweets that are incoming to Twitter [11].

Another challenge faced was to collect user location so that the distance metric may be computed. Twitter offers the option to embed geo-tagging in tweets, which contains the latitude and longitude of the tweeter when posting the tweet. Using the filtering feature of the Streaming API, it is possible to have Twitter only forward tweets with embedded geolocation information.

As a note, there was a rate-limiting bug in the Streaming API that was present at the time of data collection. Further detail is given in the Appendix.

2.2 Database Setup

Once the method for collecting a random sample of Twitter's communication was resolved, the next issue was how to organize and store all of the data related to Mini-Twitter. Presumably, over a week's time, the amount of messages collected from Twitter would be quite large (~30 million messages). In order to efficiently manage this data, the database management system MySQL was used [15]. A Java application served as the connector to the Twitter Streaming API. The Java library Twitter4J [16] offered an intuitive interface for establishing a filtered connection. Tweets were then stored on a MySQL server through the use of JDBC [17]. The MySQL database "Mini-Twitter" contains two tables, "Nodes" and "Links." The creation of these tables were as follows:

1. For each tweet, save the tweeter, their number of followers, and their location (latitude and longitude, embedded in the tweet) as an entry in the Nodes table.
2. For each mention (@sn) in the tweet, store the tweeter and the mentioned user (along with the time of the tweet) as a tuple in the Links table. Do the same for any reply.

One glaring flaw in this database construction is the fact that there is no guarantee that a user mentioned will ever be caught and placed in the Nodes table. This is due to a number of reasons, the most obvious of which is that not all users have geolocation tagging enabled. As a result, some preprocessing must be done on the Links table before Mini-Twitter can be analyzed.

The Mini-Twitter database can be interpreted as a graph, with each entry in Nodes representing a unique user, and each entry in Links representing a unique directed edge from the sender to the receiver. In this graph, the nodes reside on a globe, and all links span around the globe. The physical location of each node on the globe bears significance, as it indicates the physical location of the user on Earth.

2.3 Distance Metric

There were a couple challenges that came up when attempting to calculate the true physical distance between two users. The main issue arose due to the fact that, for each user, the only location information available from Twitter is the latitude and longitude of the user at the time of posting the tweet. This limited information lacks elevation, a crucial parameter involved in calculating the true distance between two users. The other main issue stems from the fact that the Earth is not a perfect sphere, and as such the distance between latitude/longitude pairs varies slightly depending on the location.

In order to solve these challenges and provide a distance metric to be used for later analysis, some simplifying assumptions were made. These assumptions were based on previous work in calculating distance using only latitude/longitude pairs, taken from the Great Circle Algorithm. The Great Circle Algorithm essentially treats the Earth as a perfect sphere, and calculates distance between points on that sphere using the Haversine Formula, giving an "as the crow flies" distance [19]. This assumption ignores elevation as well as subtle nuances in the Earth's shape. From this assumption, we can create a score for distance between two users as:

$$\text{distance(usr1, usr2)}$$

$$= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

where Φ_i is the latitude of usr_i, λ_i is the longitude of usr_i, and r is the radius of the Earth (mean radius r = 3958.75 miles [14]). Testing showed that this metric worked reasonably well for estimating true distance, and was never off by more than a factor of two.

3. DATA COLLECTION

Tweets were collected over the course of one week, from 5/3/2014 02:40 to 5/10/2014 13:03. Figure 1 shows two histograms of the number of tweets collected, binned by hour. The larger column graph shows the total number of tweets collected; the smaller column graph shows the number of usable tweets after preprocessing.

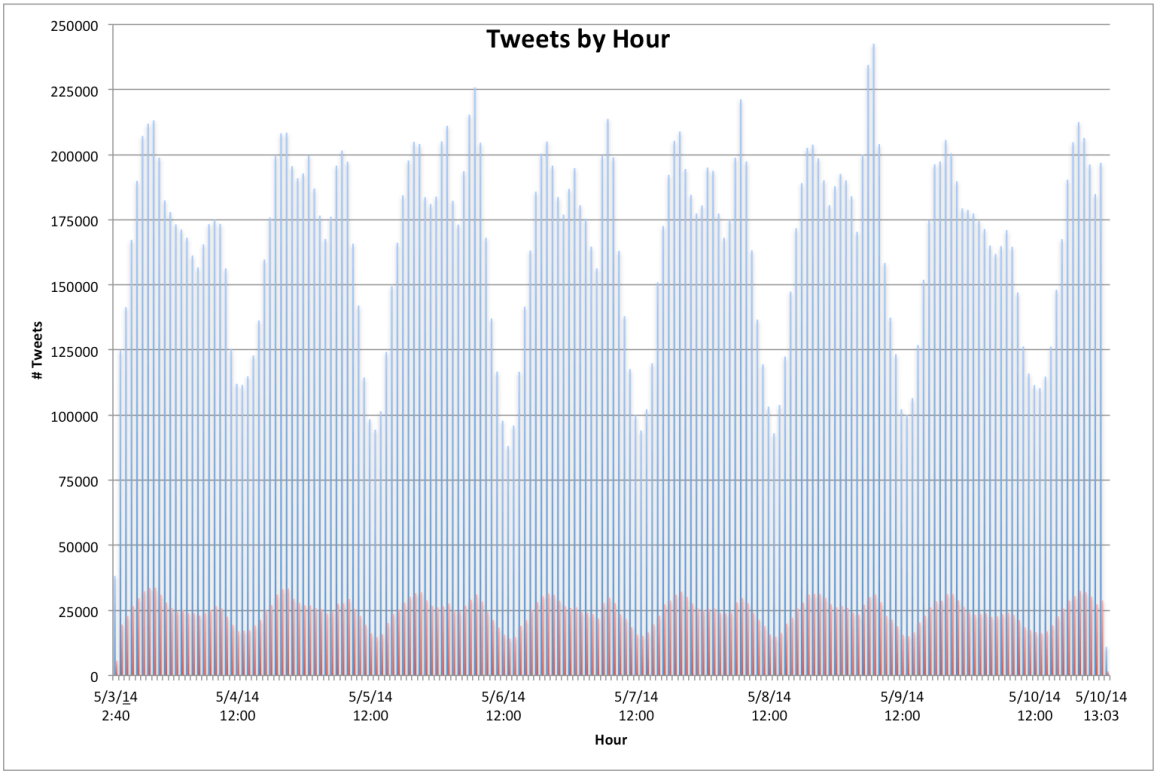


Figure 1: Number of Total Tweets by Hour and Useful Tweets by Hour - PST

For each link in our Links table, if the tweeted user was never added to our Nodes table, that link is dropped. For those familiar with web crawlers, this is comparable to a restricted web crawler that explores only outbound links that point to a trusted domain. In this case, the trusted domains are in the Nodes table, and all other outbound links are ignored. Overall, around 4.5 million usable links were collected. Peak traffic time was around 8-10AM and 6-9PM PST.

Figure 2 shows a histogram of the number of users added to the Nodes table, binned by hour. One interesting thing to note from this figure is that the number of users collected per hour does not diminish by much over the course of the week. This indicates that there are a large number of compulsory misses. One conclusion to make from this is that, over the course of the week, we have only explored a very small fraction of the total number of users on Twitter. Thus, while our database is a valid random sample, it is very small sample with respect to all communication.

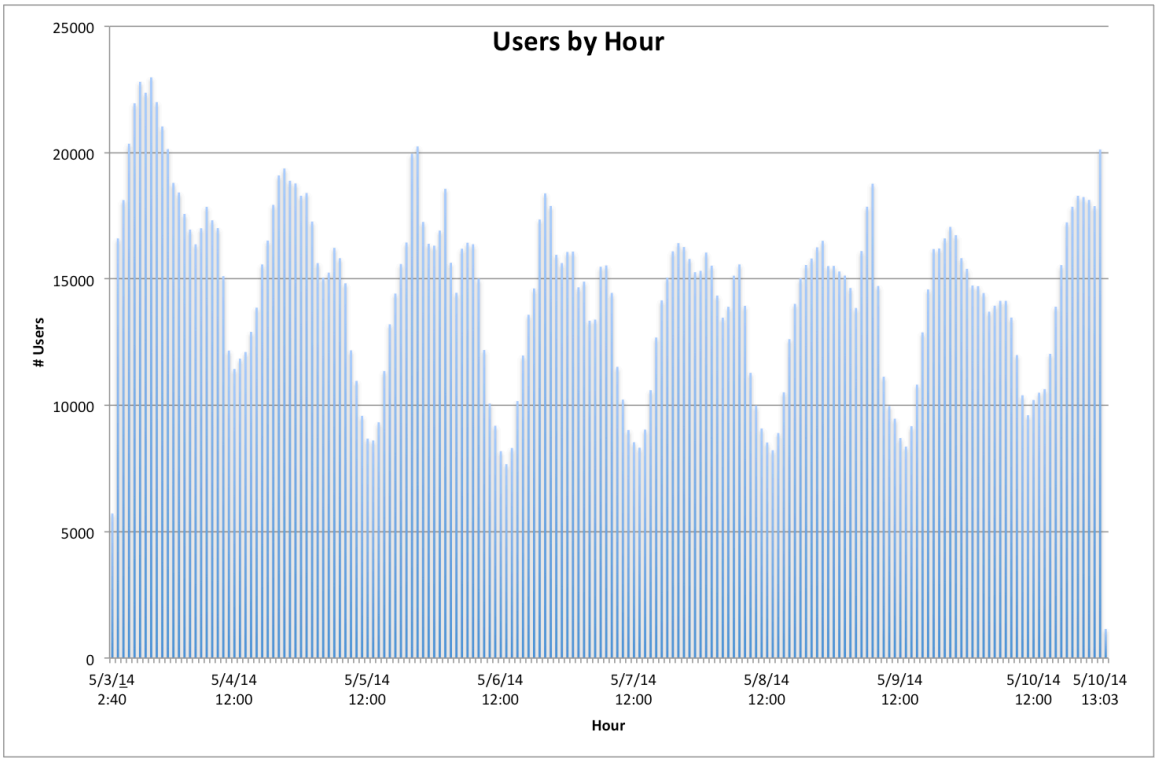


Figure 2: Number of Users Collected by Hour - PST

4. IDENTIFYING TRENDS

With the Mini-Twitter database providing a random sample of Twitter users' communication, the challenge of identifying patterns and trends is reduced to simply discovering characteristics within Mini-Twitter. Before proceeding with this analysis, the assumptions that were initially made were as follows:

4.1 Assumptions

1. The Twitter Streaming API is providing truly random tweets.
2. This week of data collection is representative of a normal week.
3. There is no correlation between communication behavior for users with private profiles or hidden tweets.
4. There is no correlation between users with embedded geolocation turned on or off.
5. Users are mostly stationary; that is, if a user tweets from a location today, he will very likely tweet from nearby tomorrow.

All of these assumptions are reasonable for the most part. The only one that could be considered questionable is the last assumption, regarding user movement. In order to test this assumption, an experiment was performed on user locations. For six days, all user locations were collected. Then, after three days, user locations were again collected for six days. For all of the users that were observed in both collections, their movement was

calculated. Figure 3 shows a normalized histogram of these users based on the distance that they moved.

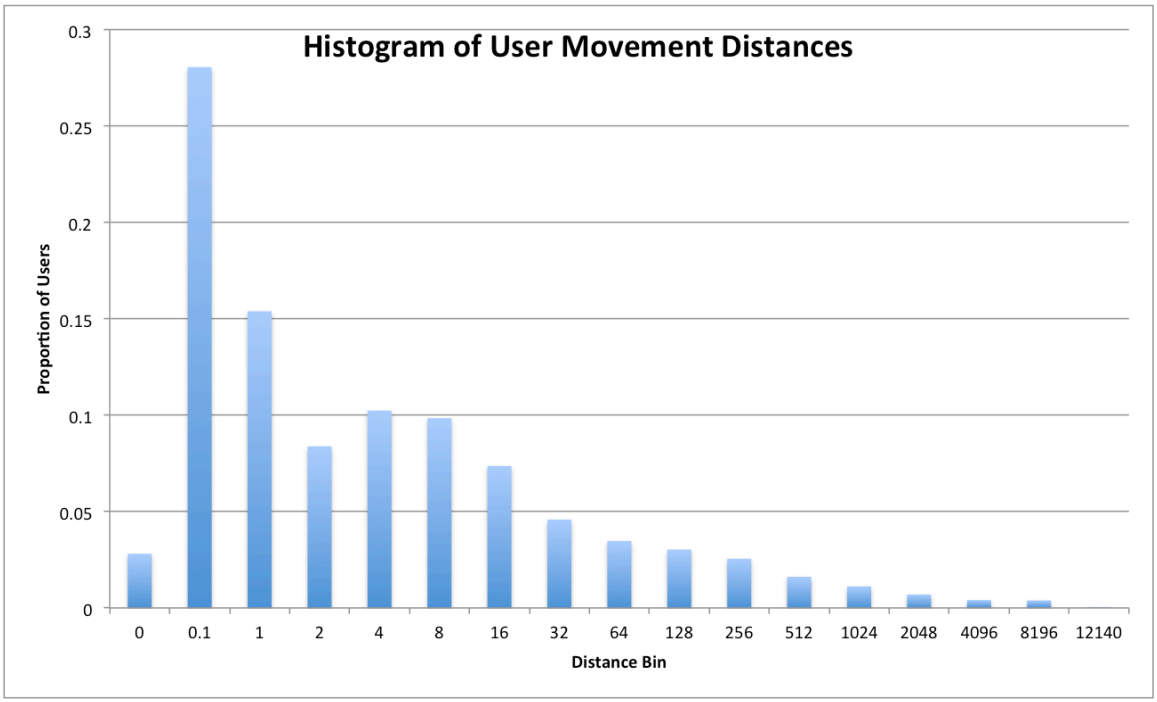


Figure 3: User Distance Change (miles)

From Figure 3, we see that ~30% of users stayed within 0.1 miles from their first location, and ~90% stayed within 50 miles. From these results, it is clear that while most users stay in the same location, a good portion of users travel very large distances, even within a week. Thus, the fifth assumption must be retracted. This is still acceptable, since the random sampling will catch a proportion of users on vacation that is fairly similar to the actual proportion of users that are on vacation (or away from home) on average.

4.2 Link Distance CDFs

Now that the assumptions have been clarified, Figure 4 shows a CDF of link distance for all tweets collected. Initial observations meet expected results, in that most communication is very local (e.g. 50% of all communication < 16 miles). Upon further inspection, the distribution seems fairly regular. When fitted in MATLAB, the CDF matches fairly well with the lognormal distribution [18]. Figures 5 & 6 show the best-fit lognormal curve overlaid on top of the actual CDF (μ=3.05, σ=2.88).

The only significant deviation from the lognormal graph occurs as we approach the max distance. This is due to the fact that a lognormal curve is asymptotic as the potential distance approaches infinity. Of course, this is not realistic, since in reality the furthest communication link is from one end of the Earth to the other. However, aside from this small deviation, it is reasonable to conclude that distance in communication can be modeled using a lognormal distribution.

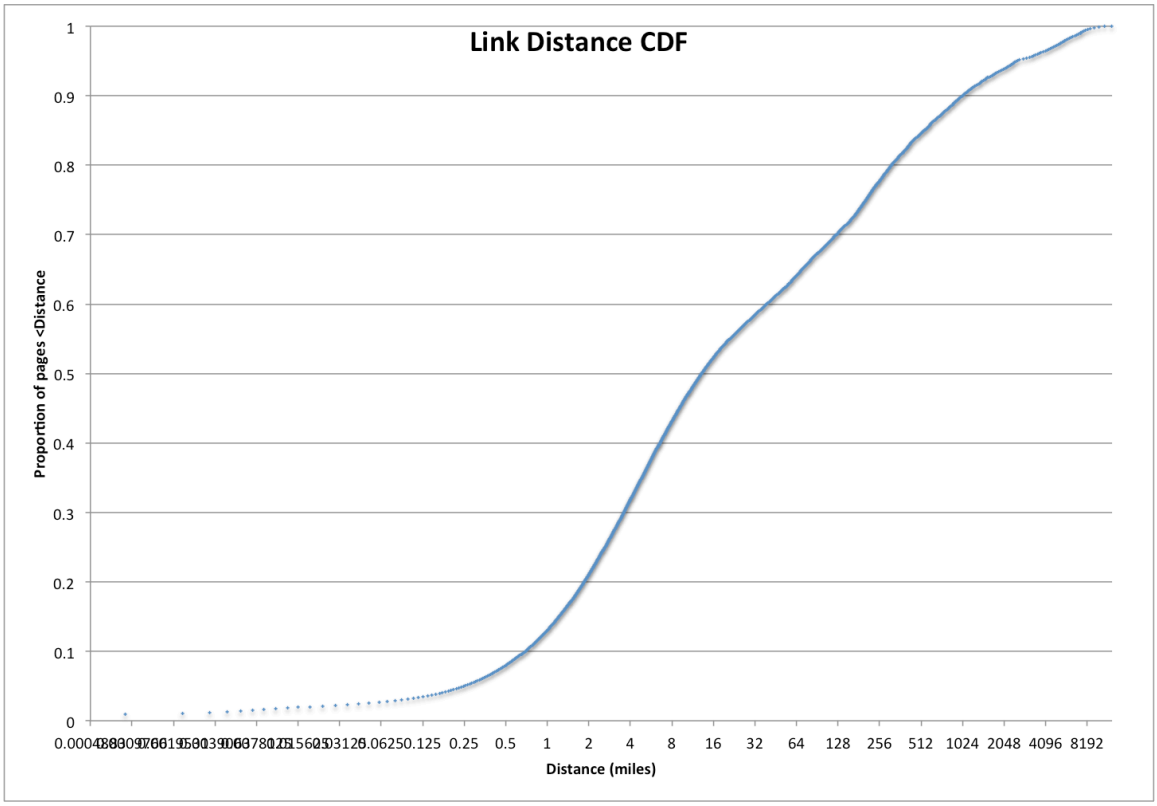


Figure 4: Link Distance CDF for all links

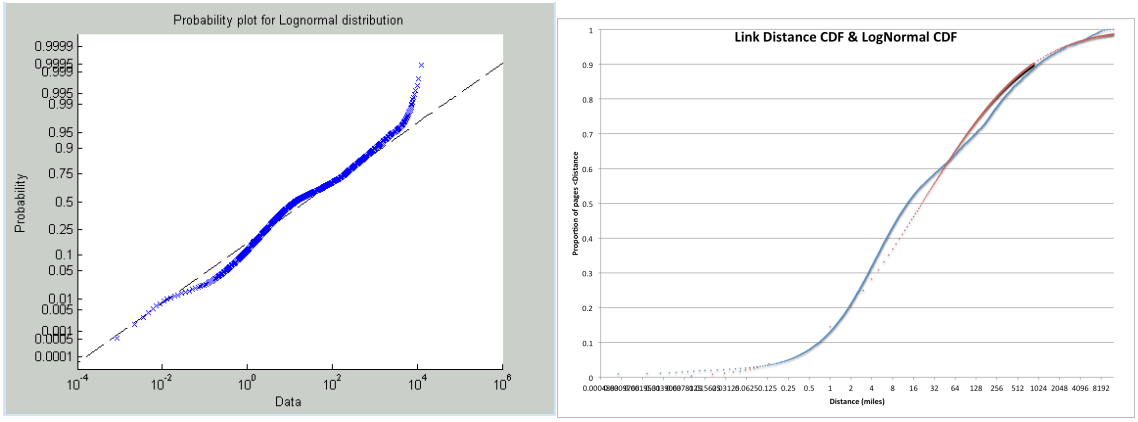


Figure 5 & 6: Best Fit Lognormal and Actual CDF (log-log scale and log scale)

While the global behavior proves to be almost identical to a lognormal distribution, this model breaks down when considering links based on location. Figure 7 shows the link distance CDF broken down by location, where only links originating from each country/city are considered.

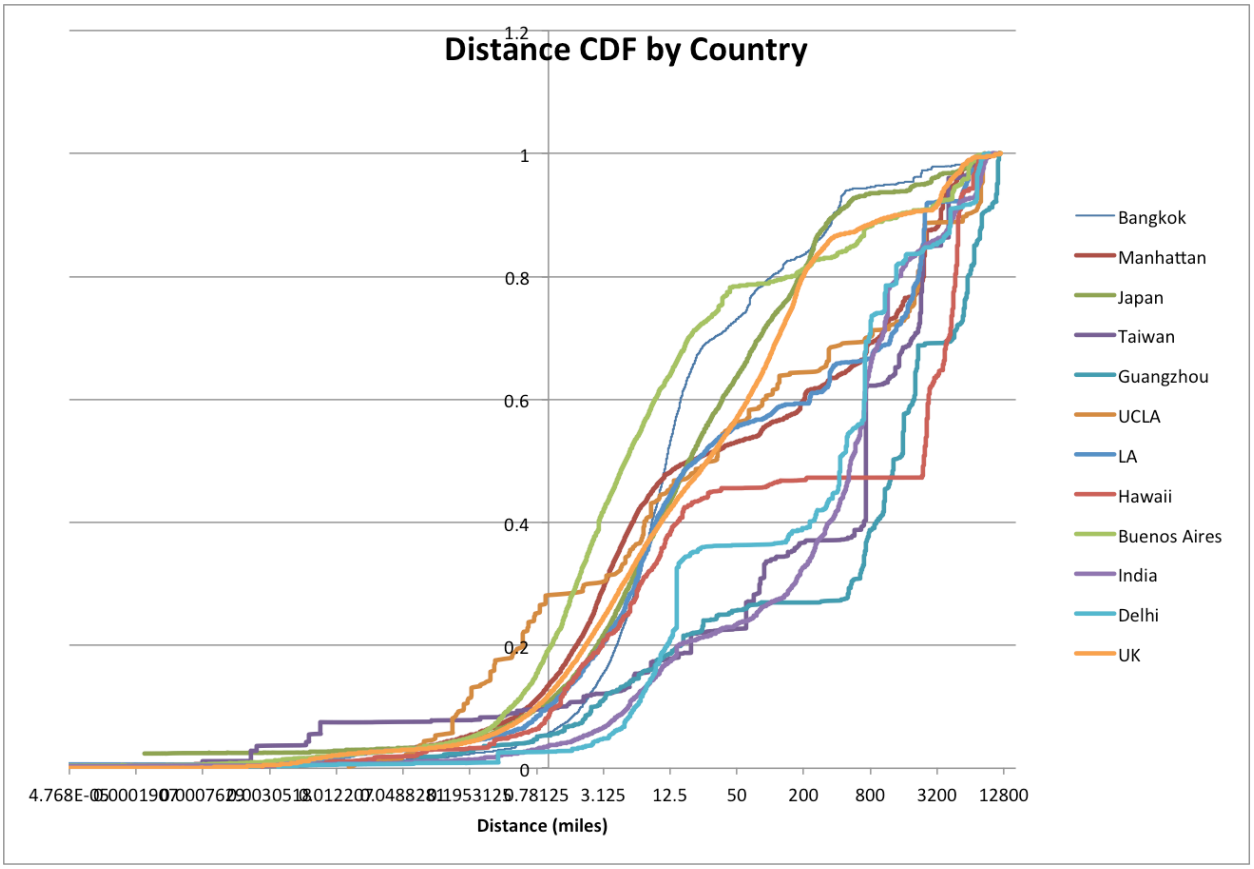


Figure 7: Link Distance CDF by Country

The most notable example of this model breakdown is with Hawaii. From Hawaii's CDF curve, approximately 50% of communication falls within 12 miles, while the other 50% of communication has a range greater than 2000 miles. Of course, this makes sense, since Hawaii is isolated in the middle of the Pacific Ocean, so communication must either be within Hawaii or far enough to span the Pacific Ocean. Thus, geographic characteristics will affect the ability for users to communicate over certain distances.

When modeled with a best-fit lognormal curve and 95% confidence intervals, many CDFs ended up falling into one of three categories: mostly short-range (Bangkok, Japan, Buenos Aires, UK), mostly long-range (India, Delhi, Guangzhou, Taiwan), or a hybrid of the two (UCLA, LA, Manhattan). These results clearly indicate that some countries are less willing to make international and far-reaching connections, while others spend the majority of their communication sending messages to users that are thousands of miles away.

Next, users were partitioned based on their follower count, and a link distance CDF was constructed for each group. Figure 8 shows the CDFs when partitioned by outbound links (i.e. based on tweeter's number of followers), while Figure 9 shows the CDFs when partitioned by inbound links (i.e. based on tweeted number of followers).

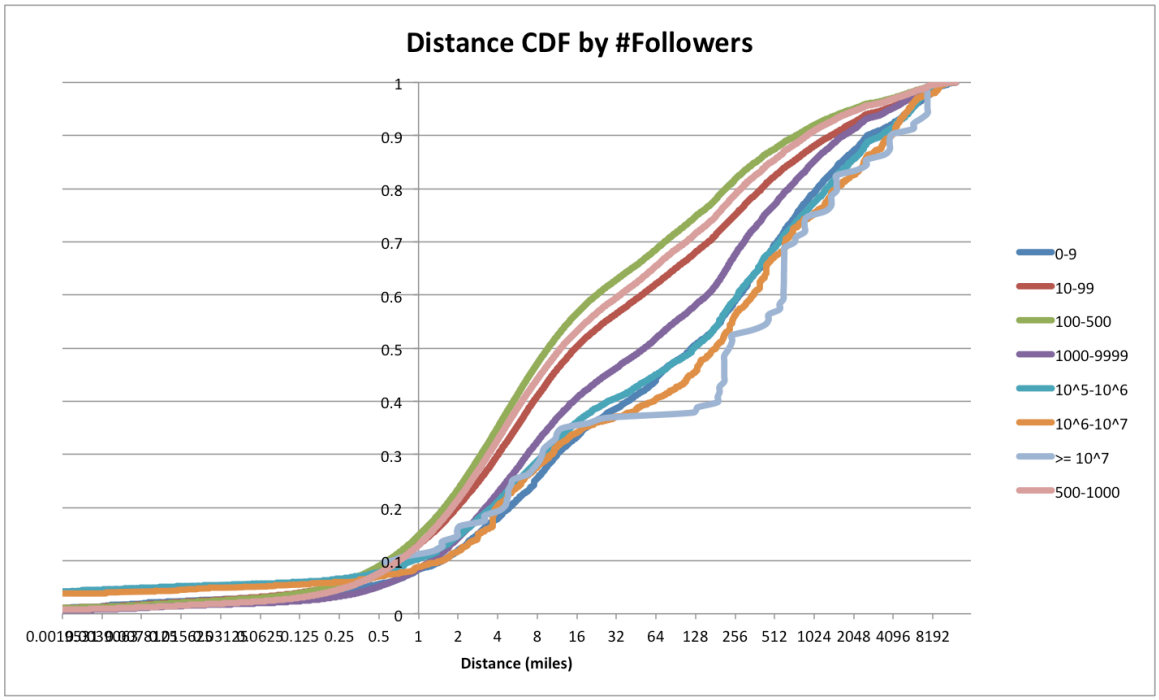


Figure 8: Link Distance CDF by Follower Count (Outbound Links)

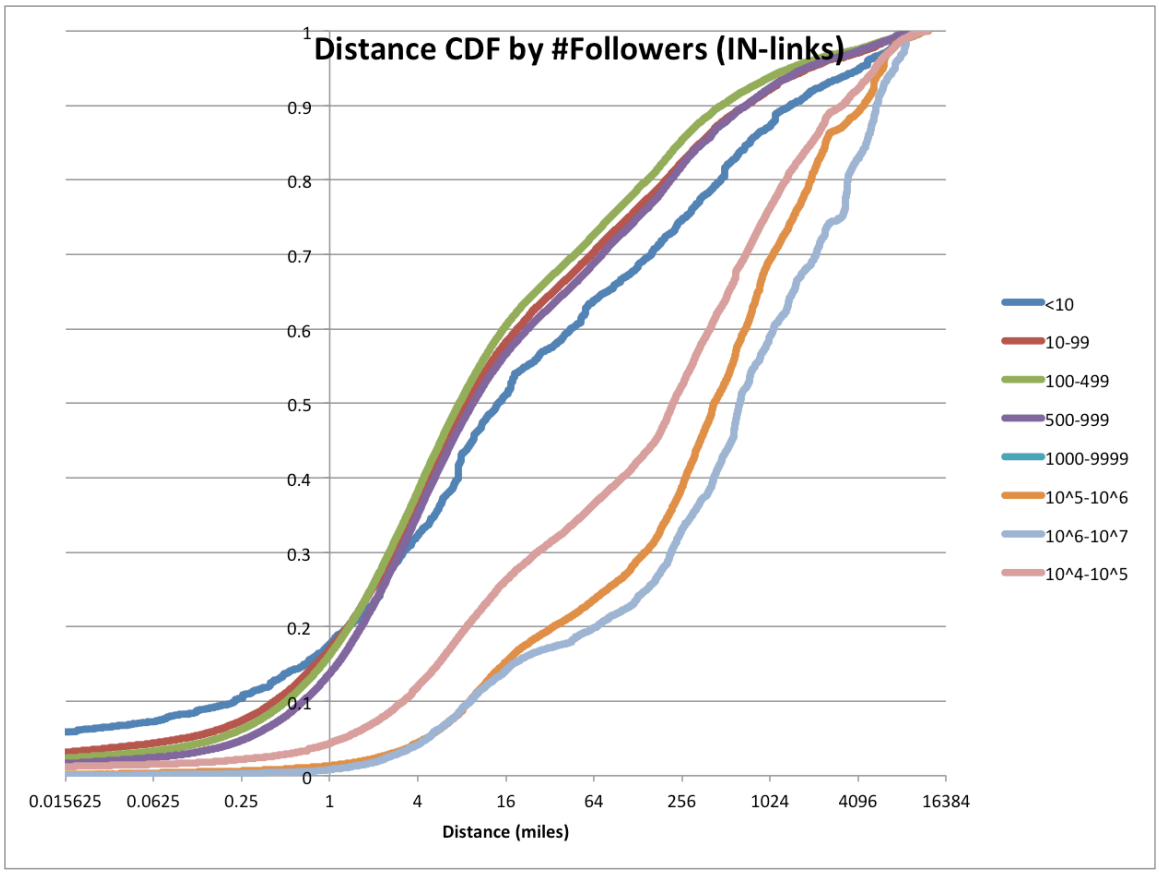
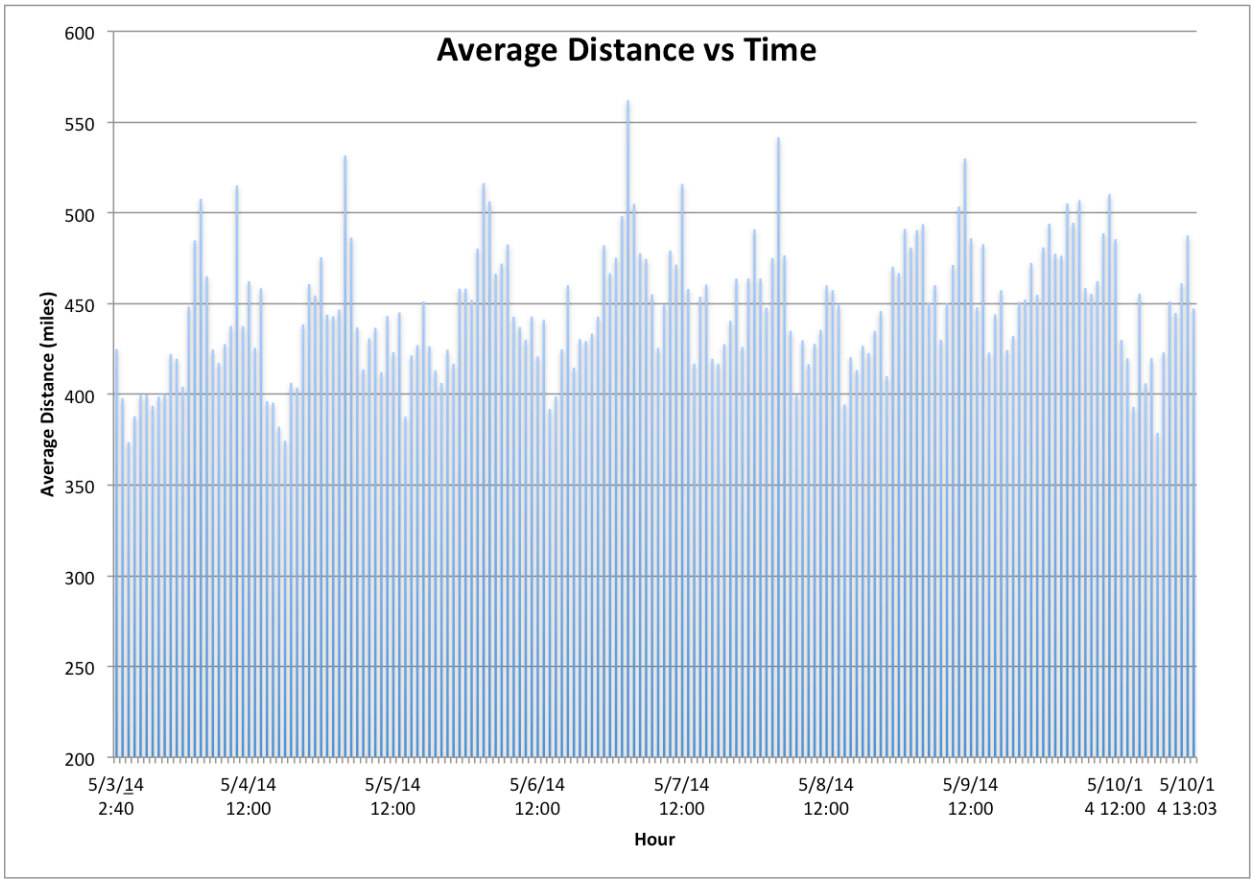


Figure 9: Link Distance CDF by Follower Count (Inbound Links)

There is clearly a significant difference in communication behavior between those with a large number of followers versus a smaller number of followers. In fact, in the range of 100 – 10,000,000 followers, the link distance CDF move monotonically to the right, indicating that individuals with more followers are more willing to tweet from far distances, and be tweeted to from far distances. This follows natural intuition, since those that are more popular are more likely to travel and form connections to those that are not physically nearby. Confidence interval testing indicated statistical significance over a difference of two orders of magnitude or more.

Finally, links are grouped by hour in order to observe if there is a correlation between the time of day and the distance of communication. Figure 10 shows the average distance of links for each hour over the span of the week. There is definitely a noticeable trend indicating that some sort of correlation exists. Peaks and troughs in the graph seem to occur regularly and daily, with the peak distance traffic occurring around 3pm PST. It is interesting to note that peak traffic time (6-9pm) is different from peak distance traffic (~3pm).



Overall, there are significant trends in the behavior of user communication with respect to distance. Whether partitioned by location, user characteristics, or time, there are clear indications that different groups of users have significantly different communication habits.

4.3 Observations

The fact that the distribution of communication follows a lognormal distribution so closely is quite surprising. It provides marketing and social scientists an idea of the "reach" of viral trends in social networking; that is, what other areas will be influenced when a certain location or demographic is targeted. This becomes especially important when considering specific locations, since the communication habits vary significantly with location. Some countries, such as Japan and the UK, keep most communication local. Some factors that affect this may be language and the smaller size of these countries. Nevertheless, the communication patterns are very different from countries like India and China, who have over half of their communication spanning ranges greater than a thousand miles. This may be due in part to the larger size of the country, but must also be due to the emigrate nature of these countries. Of course, the ability to have half of total communication span over a thousand miles, especially in third world countries such as India, is something that was made possible only by the rise of the Internet and low-cost online messaging. Taking advantage of the potential of effective long-range communication is definitely an explanatory factor in these countries' emerging economies.

It is important to note the fact that only 40% of the world population currently has access to the Internet. The lack of access for the 60% of disconnected individuals affects their ability to expand communication and observe the economic benefits of interconnectedness, following a "rich get richer" predicament. In the future, as the price of bandwidth decreases and infrastructure improves, the barrier of physical distance can finally and truly be broken.

The time of day also affects user communication. Users are more willing to communicate long distances at certain hours (~3pm PST), and shorter distances at other hours (~3am PST). The fact that peak distance traffic occurs at different hours than total Twitter traffic has implications in data center optimization, in that resources may be specially allocated for long-range messages during peak distance traffic (~3pm PST), and then reallocated to handle normal traffic during Twitter's normal peak traffic (~6pm PST).

There were a number of bots within our dataset; certain users displayed communication patterns that are infeasible. Two examples of such are users isomorphisms and equivocagent, the first of which consistently tweets from the tip of the south pole, while the latter tweets from a random location on the globe for every new tweet.

Despite the fact that over 35 million tuples were stored in Mini-Twitter, the disk space required to store all of this was surprisingly small. Only around 6GB was necessary to store all 35 million tuples. However, even with 30 million links collected, the size of Mini-Twitter pales in comparison to all of the communication that is taking place. It would be advantageous to expand Mini-Twitter with more users and data in order to reduce noise in modeling and identifying characteristics in locality distributions.

5. FUTURE WORK

In the future, it would be beneficial to collect users and tweets for more than simply one week. Ideally, Mini-Twitter would continue obtaining data until the number of users

collected begins to approach the total number of Twitter users. First and foremost, this would lead to more confident results. However, another benefit of having a more robust dataset is that the actual graph structure of social networking can be observed, with respect to strongly connected components, in components, out components, in-degree and out-degree. Initial testing showed the in/out degree to follow power law distributions, but there is not enough evidence to confirm this trend.

5. CONCLUSION

To conclude, the initial question of understanding user communication due to the rise of the Internet was answered at both a global and local level. There are obvious communication characterizations present in social networking that have implications in many fields, especially marketing and the social sciences. There exist significant patterns in communication with respect to physical distance, and these patterns vary significantly by the group of users observed. Overall, the global trend in communication follows a fairly lognormal distribution, with the median link distance of around 15 miles. However, this model breaks down at a sub-continental level, due to both geographic and cultural differences. When categorizing users by their popularity, it was observed that more popular users span farther distances, in terms of both outbound and inbound links.

Finally, the original assumption that users are stationary was shown to be invalid, and a distribution of the change in location was given. While this does not affect our results, it again provides key insight into human behavior.

A study of this nature has never been conducted before. Some of the results observed followed natural intuition, while others proved surprising. Overall, these conclusions indicate that, while individuals are still most likely to communicate nearby, the rise of the Information Age has allowed for individuals to more readily communicate with the rest of the world. In fact, some countries have over half of their online communication spanning great distances, something that only modern computing technology could provide for. As mankind progresses further into the future, it stands to reason that the world will become more and more interconnected, leading to a true embodiment of the saying “it’s a small world after all.”

REFERENCES

[1] Humbert, Mathias. "Technology and Workforce: Comparison between the Information Revolution and the Industrial Revolution." *University of California, Berkeley School of Information. p* 5 (2007).

[2] Tam, C. M. "Use of the Internet to enhance construction communication: total information transfer system." *International Journal of Project Management* 17.2 (1999): 107-111.

[3] St. Aubin, Emma. "We Are the Social Media Generation." *We Are the Social Media Generation*. The Pointer, 2011. Web. 18 Apr. 2014.

[4] Dekker, Anthony. "Conceptual distance in social network analysis." *Journal of social structure* 6.3 (2005): 31.

[5] Baum, Leonard E., and Melvin Katz. "Convergence rates in the law of large numbers." *Transactions of the American Mathematical Society* 120.1 (1965): 108-123.

[6] https://www.facebook.com/

[7] Tam, Donna. "Facebook by the Numbers: 1.06 Billion Monthly Active Users - CNET." *CNET*. CNET News, n.d. Web. 3 June 2014. <http://www.cnet.com/news/facebook-by-the-numbers-1-06-billion-monthly-active-users/>.

[8] https://www.facebook.com/note.php?note_id=91351698919

[9] https://twitter.com/

[10] https://dev.twitter.com/docs/api/streaming

[11] "GET Statuses/sample." *Twitter Developers Website*. Twitter, n.d. Web. 10 May 2014. <https://dev.twitter.com/docs/api/1.1/get/statuses/sample>.

[12] "Sudden Increase in 420 (Exceeded Connection Limit for User) Responses from Streaming API." *Twitter Developers Website*. Twitter Discussions, n.d. Web. 2 May 2014. <https://dev.twitter.com/discussions/27717>.

[13] https://plus.google.com

[14] "Great Circle Algorithms." *OpenLayers*. Trac, n.d. Web. 22 May 2014. <http://trac.osgeo.org/openlayers/wiki/GreatCircleAlgorithms>.

[15] MySQL, A. B. "MySQL." (2001).

[16] Yamamoto, Yusuke. "Twitter4J-A Java Library for the Twitter API." (2010).

[17] Bales, Donald. *Java programming with Oracle JDBC*. O'Reilly Media, Inc., 2002.

[18] Guide, MATLAB User'S. "The mathworks." *Inc., Natick, MA* 5 (1998).

[19] Veness, Chris. "Calculate Distance, Bearing and More between Latitude/Longitude Points." *Movable Type Scripts*. Movable Type Scripts, n.d. Web. 2 May 2014. <http://www.movable-type.co.uk/scripts/latlong.html>.

APPENDIX – Rate Limiting Bug

While gathering tweets, there was a fascinating bug that severely hindered progress. To preface this, Twitter's online API is very restrictive, with their REST API limiting calls to one per minute. This is a very frustrating limitation, so to overcome this we chose to use the Streaming API.

On April 15, 2014, Twitter released an update to their Streaming API. The update included a number of new features, most notably that applications must use double authentication (developer and application) in order to access the API. There was a bug in their release that caused users connecting to a stream to continuously receive "Connection Limit Exceeded" errors. This bug is described at [12]. An error would display indicating that the user was "Running too many copies of the same application authenticating with the same account name," despite the fact that I only had one instance of my Java application running at a time.

Something interesting that was noted when observing this error was that the application could easily connect the first time, however subsequent re-connections would fail. After posting on the discussion forums at [12] and exchanging emails with a twitter developer, we discovered that the bug was caused by a failure to reset the connection count when the stream was disconnected in a non-ideal way (i.e. ctrl-c instead of stream.close()).

After this bug was identified and fixed, connecting to the stream was much simpler and easier, mitigating one of the limiting aspects of using this particular API. The exact message output due to this bug was as follows:

"Too many login attempts in a short period of time.

Running too many copies of the same application authenticating with the same account name.

Easy there, Turbo. Too many requests recently. Enhance your calm."