A Statistical Analysis of a Time Series of Twitter Graphs

David J. Marchette*

Abstract

In this paper I describe a set of Twitter data that we have been collecting for nearly two years. Using the Twitter streaming API, we collect all tweets geo-located within a set of rectangles covering the main land-masses of the world, as well as tweets containing certain key phrases. We collect "all" geo-located tweets, in the sense that Twitter provides all the tweets that are geo-located within the rectangle, provided the volume does not exceed a fixed limit. These tweets define a "mentions" digraph – each user id is a vertex and there is an edge from s to t if a tweet from s mentions t: @s:"@t u wanna go to lunch?". These mentions digraphs can be computed on time intervals to produce a time series of graphs. These graphs tend to have power law degree distributions, and I will describe the graphs and discuss some thoughts on how one might model these graphs. Using the graphs, I will discuss methods for inferring node attributes, such as the geo-position of a user whose tweet is not geo-located, or detecting spoofed geo-locations.

Key Words: Social media, twitter, geographic inference, analysis of large graphs

1. Introduction

Twitter is a service that provides easy access to microblogs by millions of users. Users can send out short messages, which are seen by anyone who chooses to follow the user, as well as anyone who chooses to access the Twitter API[1]. Each microblog, or *tweet* consists of up to 140 characters, and can contain, in addition to the content, one or more links to other content on the web, and references to users. If a user is referred to (mentioned) in a tweet, that tweet is highlighted to the user. Thus the tweet can be thought of as being set to the referenced user(s), even though it is also seen be all followers of the sender.

Tweets are analyzed for several reasons. There is considerable interest in using them for public health (Corley et al. [2010], Lampos et al. [2010], Aramaki et al. [2011], Collier et al. [2011], Dredze [2012]), for understanding the spread of information in social networks (Huberman et al. [2008], Lerman and Ghosh [2010], and detecting, tracking and analyzing social unrest (see the Special Issue Title: New Media and Social Unrest, of the American Behavioral Scientist, `http://abs.sagepub.com/content/57/7.toc`). The ease of access to Twitter information makes it a valuable tool for understanding social media, social interactions, and social networks.

There are several graphs that one might investigate that describe the social network of Twitter users. The friends/followers graph shows who is following whom, in much the same way that such mechanisms are implemented in other social media such as Facebook. The graphs that we will consider in this work are the mentions graphs, with an edge between each user and all the users that are mentioned (referred to) by the user. A typical such tweet (made up for purposes of illustration) is:

```
Sillyness142:
@Python4Ever: chk out the news re parrots http://t.o/Ssaw2l
```

*Naval Surface Warfare Center, 18444 Frontage Rd, Suite 327, Dahlgren, VA 22448. This work funded in part by the NSWC In-House Laboratory Independent Research (ILIR) program. NSWCDD-PN-15-00076 Distribution Statement A: Approved for Public Release. Distribution is Unlimited.

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

Tweets like this indicate more than a shared interest. They indicate at least some level of personal interaction and in many cases indicate that the two users know each other personally. This is a better indication of friendship than the friends/followers graph, since one may follow many people one doesn't know, but tweeting **to** someone indicates a certain amount of familiarity (although as with any interaction on the Web, this shouldn't be taken too literally).

The graphs we will be concerned with in this study come primarily from the United States: a rectangle covering the lower 48 states was defined, and the Twitter API was queried to return all tweets with a geographic location within the rectangle. The API will return all tweets matching a query, up to an upper limit. Experiments have shown that while we do hit the limit during some peak times, we do collect nearly all relevant tweets.

We will also consider a single graph constructed from similar rectangles placed around all the populated continents. This typically results in about 10 million tweets per day. In this larger collection we believe we do hit the limits of the API fairly regularly, but have not explored this further.

Not all tweets have a geographic location (estimates have ranged from 0.2% to 3% of tweets have a location – see Leetaru et al. [2013]). The location is typically set by the device used to send the tweet. For example, a geo-enabled cell phone will give the GPS coordinates of the phone.

Not all geographic locations are true. It is easy to set (or spoof) the geographic location of a tweet, and this is sometimes done for privacy reasons, or simply for amusement. It is also sometimes done by researchers studying Twitter (and is how we determined that we collect most geo-located tweets). It can also be set by a web page. Many pages have a Twitter icon that allows the user to tweet out a link to the page, and some of these pages will set a geographic location on the tweet (for example, a news story may set the location to be the location of the story).

The mentions graph is a dynamic graph, potentially adding an edge every time someone tweets. We will consider static versions of this graph. For a given time period T, define the mentions graph G_T to be the directed graph with edges:

$$u_1 \to u2$$

whenever user u_1 mentions user u_2 in a tweet sent during T. We could keep the number of times a mention occurs, but for this study we will not.

Note that our graph has a built-in selection bias: if user u_1 allows geo-location, but user u_2 does not, we will see any edge $u_1 \to u_2$, but will not observe any edge $u_2 \to u_1$. Further, even if u_2 provides a geo-location, if this location is outside our rectangle we will still miss the edges from u_2.

Because Twitter only provides a subset of the tweets through its public API, there is always some bias in collections that use this API. There are ways to purchase the full feed of tweets, and Twitter is making the full feed available to researchers under some conditions, but for this study we had to rely on the public API.

We will focus on the month of April, 2014, and will take for our time periods the three weeks: April 6–12, 13–19, 20–26. We will also consider the daily graphs, consisting of the tweets occurring on a given day, from April 4–29 (due to power outages we did not collect all the tweets during the first three days of April).

The paper is structured as follows. In Section 2 we will investigate the structure of these graphs, through analysis of some graph invariants. We will look at some models for these graphs in Section 3, and look at using the graphs for an inference task – estimating geo-coordinates – in Section 4. Finally we will provide a discussion of the results in Section 5.

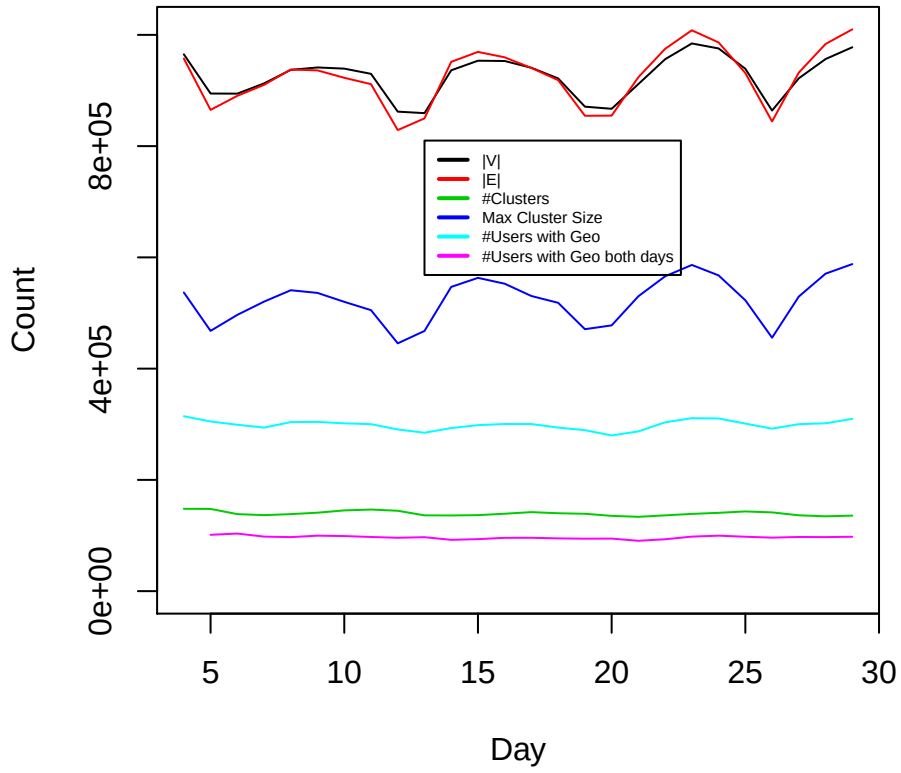

Figure 1: Graph invariants computed on the 26 daily graphs.

2. The Twitter Graphs

First consider graphs constructed each day. Each graph has vertices corresponding to the users that tweeted that day, with an edge from user u to user v if u mentions v during the day. One would expect a level of nonstationarity to these graphs, due to the fact that one would expect different behavior on the weekends than during the week, and perhaps different behavior on Mondays and Fridays than during the middle of the week. We investigate this by constructing several invariants, plotted in Figure 1.

Note the obvious seasonality on a weekly basis that is evidenced in the order of the graph (number of vertices), the size (number of edges), and the size of the largest component. The other invariants also show a very weak pattern, that might be more pronounced in a rescaling. The cyan curve corresponds to the number of vertices at day t that have at least one tweet with coordinates, and the magenta curve is the number of these that also had coordinates the previous day. These indicate that only about 32% of our vertices have coordinates, and thus that the rest are mentions of users who do not show up in the collection as tweeters, either because they chose not to tweet, or because their tweets do not contain their location.

As shown in Figure 2, the graphs have a very distinctive power-law degree distribution that is consistent across the days. The level of overplotting in this plot is quite pronounced.

Note also the large number of users with a single edge or only a few edges. This seems

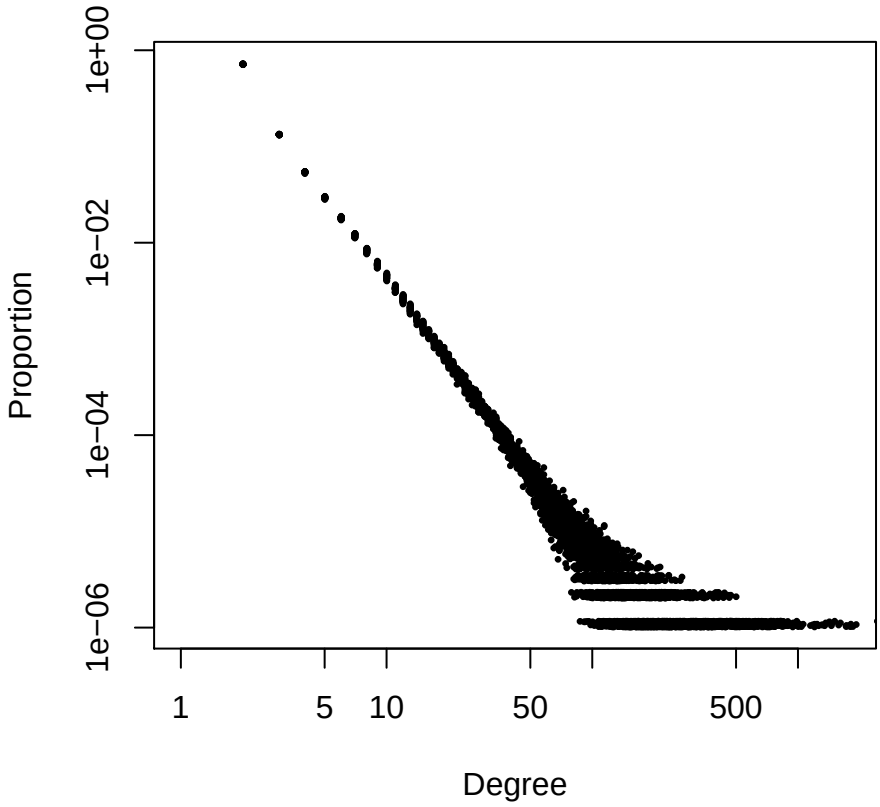

Figure 2: Degree distributions for the daily graphs. These are 26 plots of the proportion of vertices with a given degree, plotted on a log-log scale. In this plot the total degree (sum of in-degree and out-degree) is plotted.

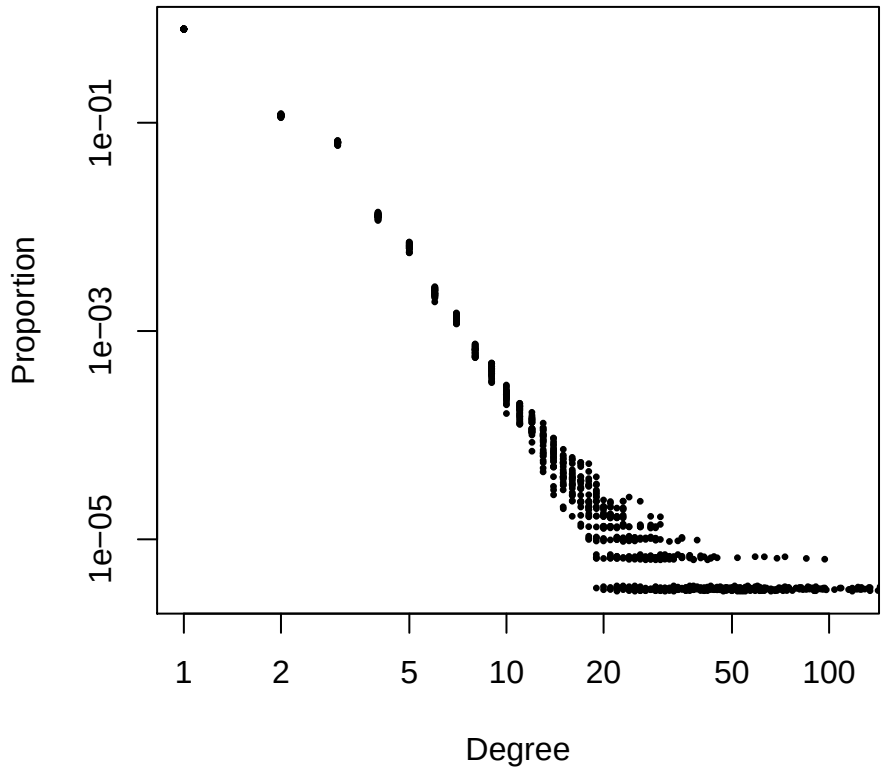

Figure 3: Degree distributions for the subgraphs of the daily graphs consisting of only those vertices that contain at least one tweet with a coordinate. These are 26 plots of the proportion of vertices with a given degree, plotted on a log-log scale. In this plot the total degree (sum of in-degree and out-degree) is plotted.

likely to be in part a result of the sampling bias, since we only observe out-edges from a fraction of the vertices. Also, since we are only looking at one day's worth of tweets, we expect to see only a few tweets per user (for most users) and hence expect many low degree vertices.

Consider the corresponding picture for the induced subgraphs consisting of only those vertices with coordinates (Figure 3). We see the same basic shape (constrained by the much smaller graphs). This combination of relatively few tweets per user (with a commensurate low number of mentions) and the obvious diurnality of the graph, leads us to consider a different time window on the data.

First, we consider graphs where the defining interval is a week. Due to sensor outages, we consider the three weeks indicated in Table 1. As can be seen in the table and the degree distributions (Figure 4), the weekly graphs are similar to, though significantly larger than, the daily graphs.

Finally, we look at a very large graph. In this case we take all the data collected in April, from all the continents, as depicted in Figure 5, and construct a graph as above. The statistics for this graph are in Table 1, and the degree distribution is depicted in Figure 4, in comparison to the other graphs. Note that the larger rectangle for North America misses

Table 1: Table of statistics for three weekly graphs constructed from the continental United States data plus the graph constructed on the world data for the month of April. The final column is the percentage of vertices in the largest cluster.

Graph	$\|V\|$	$\|E\|$	#Clusters	Max Cluster Size	% Max
April 6–12	3470349	5.301605×10^6	161969	3.024976×10^6	87.2
April 13–19	3461504	5.359283×10^6	159312	3.025135×10^6	87.4
April 20–26	3494342	5.421333×10^6	159482	3.057264×10^6	87.5
April – World	20997641	4.9128359×10^7	481061	1.9760423×10^7	94.1

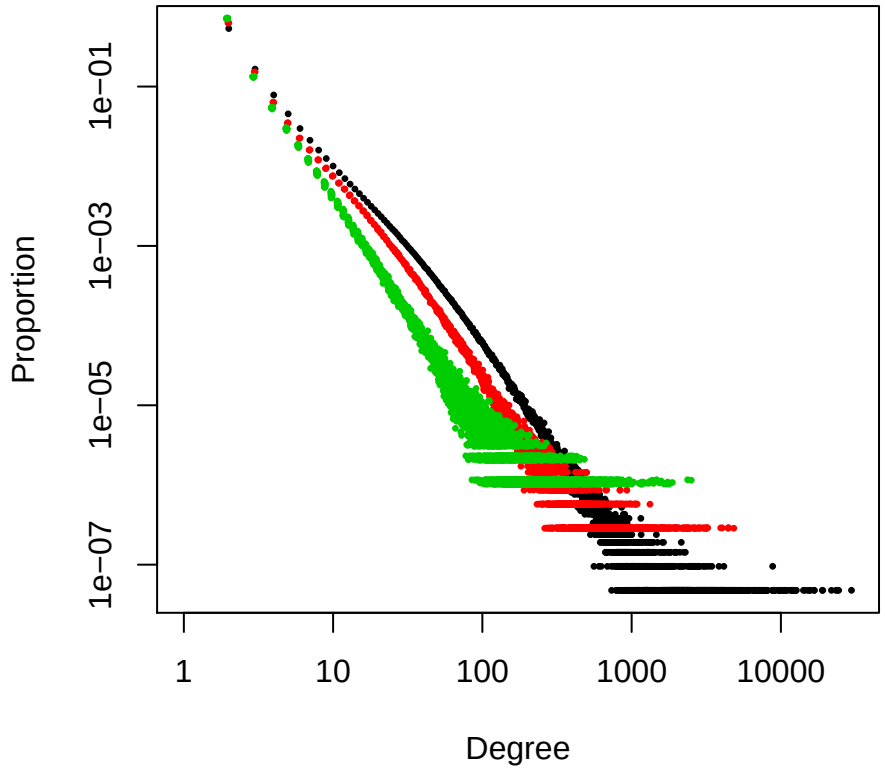

Figure 4: Degree distribution for the large graph constructed on all the data collected in April from around the world (black), the three weekly graphs (red) and the 26 daily graphs (green).

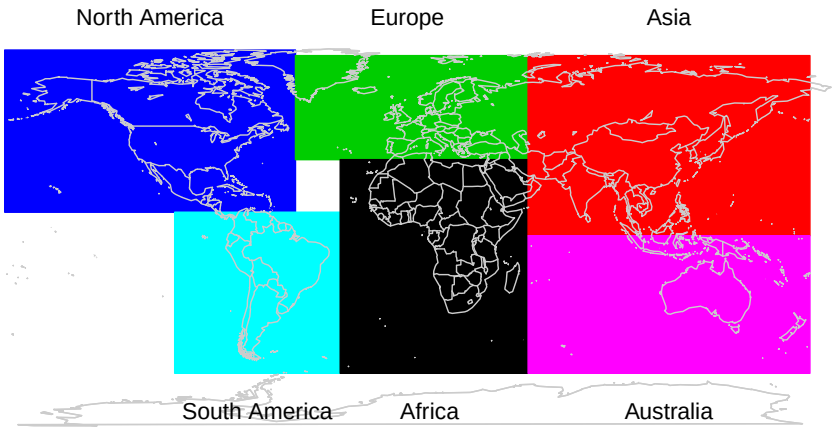

Figure 5: Rectangles defining the continental data collections. In addition to these, there is a smaller rectangle around the lower 48 states of the United States. The white regions are regions of non-coverage.

some tweets collected by the smaller rectangle covering just the lower 48 contiguous states, demonstrating that these larger rectangles can hit the limits imposed by the public API. All tweets are used to construct the graph, including those collected by the smaller US rectangle that were missed by the larger North America rectangle.

3. Graph Models

Consider a simple model for how these graphs might be generated. The random dot product model (RDPG) posits that there are two sets of latent variables for each vertex v: x^v_{in} and x^v_{out}. The probability of a directed edge $u \to v$ is the dot product of these vectors:

$$P[u \to v] = {x^u_{out}}^T x^v_{in}.$$

We can fit the model using singular value decomposition (see Marchette and Priebe [2008], Athreya et al. [2015]), however this is problematic due to the missing edges we do not observe. Since we only observe some of the edges (those from geo-located tweeters), we do the following:

1. There are 894345 vertices with 34.6% of these geo-located, so we will only generate edges from these 3.09253×10^5 vertices.

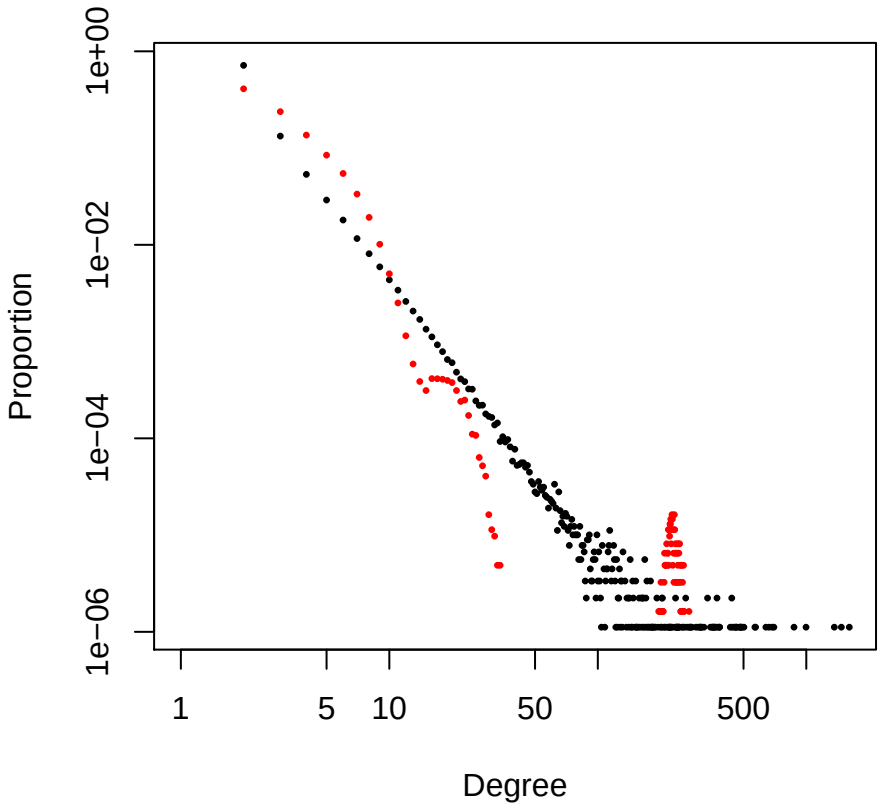

Figure 6: Degree distribution for a week graph (April 6–12) (black), compared with the random dot product model (red).

2. We believe that people "tweet locally" (as we will see in Section 4) and so we assign to each user a region. These regions are drawn from the set of US counties with probability proportional to county population.

3. We generate the in- and out-vectors according to a Dirichlet distribution that is uniform on the simplex. Each user in a county gets the same two vectors.

4. Each user tweets to the other users in their county with probability defined by that county's vectors. They also tweet (with probability 0.5) to users from another randomly chosen county with probability defined by the corresponding vectors.

5. Finally, there are 225 "celebrities", and with probability 0.1642151 a user will tweet to one of the celebrities. These values were obtained from the observed graph by defining a celebrity as one who as an in-degree greater than or equal to 100.

The degree distribution is similar, but not a perfect match. Some of this is a result of the fact that the vectors were chosen randomly, rather than fit from the observed data, and the other parameters of the model are rather arbitrary as well.

Clearly further research is needed. Essentially, we are positing a block model, where the blocks are associated with counties (plus one block for celebrities), and the between block probability is constant, the within block probability of celebrities is 0 (or rather unknown,

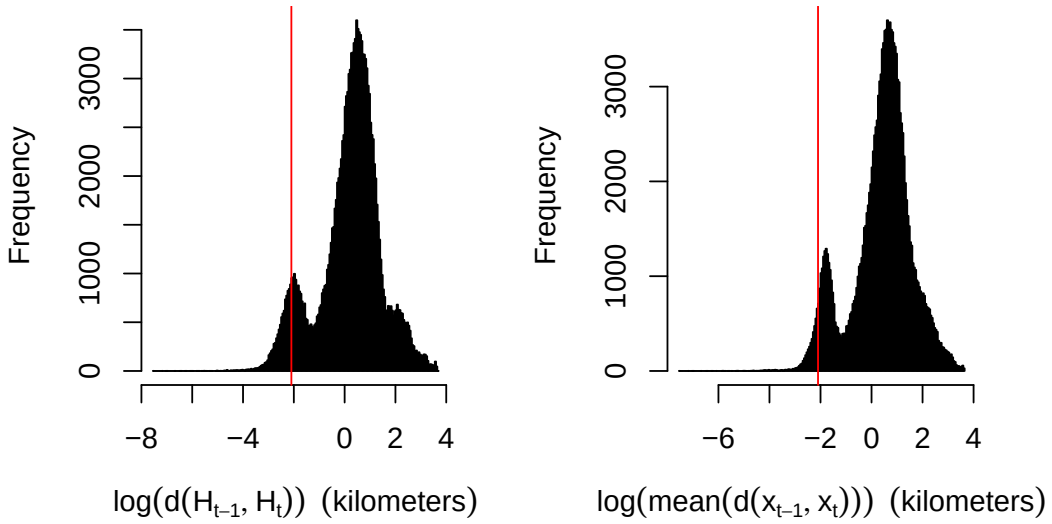

Figure 7: Distances between one day's positions and the next. The left plot shows the distance between the home position of the user on the two days, as calculated using the 2 dimensional kernel estimator. The right plot shows the mean distance between tweets on the first day versus those on the second. The red vertical line indicates 8 meters, which is approximately the accuracy of a typical commercial GPS device. There are 162182 distances represented in each histogram.

since we do not observe these edges) and the probability of an edge between one block and the celebrity block is also constant.

Methods for fitting this model, under the observational constraints, are an area of further research.

4. Geo-Inferencing

People move around, but most people stay close to home. To see this, consider Figure 7. Here we define a person's "home position" by constructing a 2D kernel estimator of their positions, and using the maximum likelihood as the position of their "home position". As can be seen by the plots, many people are stationary (or more accurately, the device they use to tweet on is stationary), with most people tweeting within about 10 kilometers of where they were the previous day. Then there are the "travelers" who traveled 100 kilometers or more from one day to the next.

People tend to tweet locally. That is, if user A tweets to user B, the odds are that the two users are geographically close. This is because one tends to tweet to close friends (close both emotionally and geographically). Of course, there are exceptions, but the data bears out the hypothesis that most of ones tweets are local.

Figure 8 illustrates this phenomenon. These are the distances between the home position of a user and the home positions of the users they mention. Note that this is somewhat biased due to our collection bias: we can only measure distance to those how report position. So, the best we can say is that if you mention someone, and both of you have turned

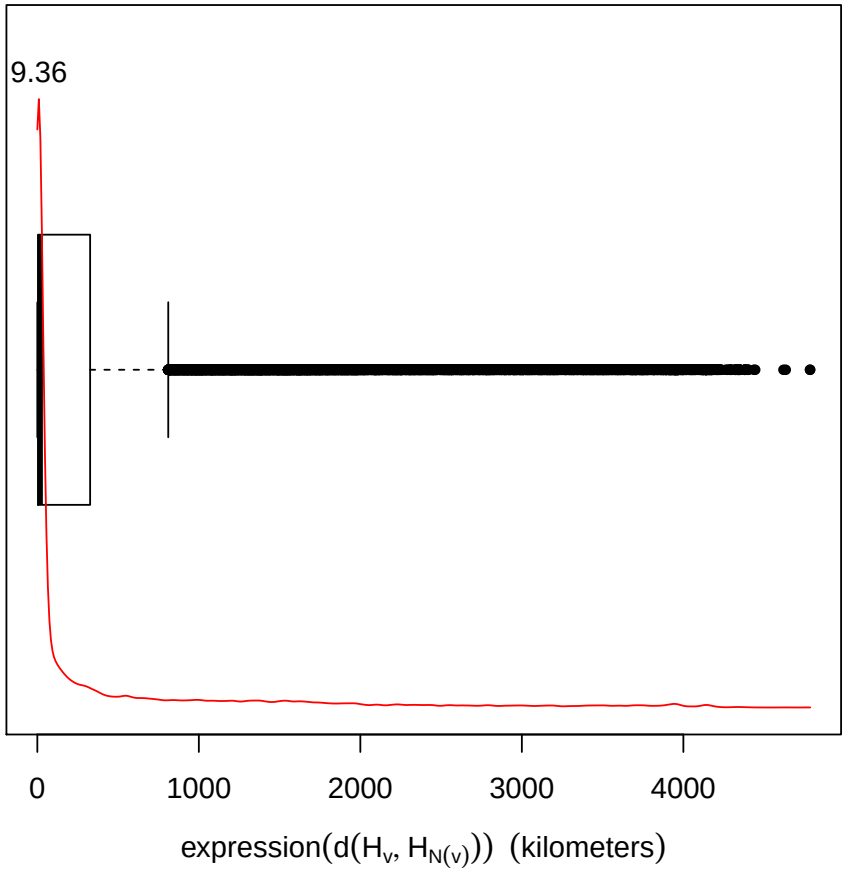

Figure 8: Distances between a user's positions and those of the people that are referenced by the user. These are the distances between home positions, as calculated using the 2 dimensional kernel estimator. The red plot shows the probability density of the distances (as estimated using a kernel estimator) with the distance associated with the maximum likelihood above the curve. There are 52396 distances represented in this plot.

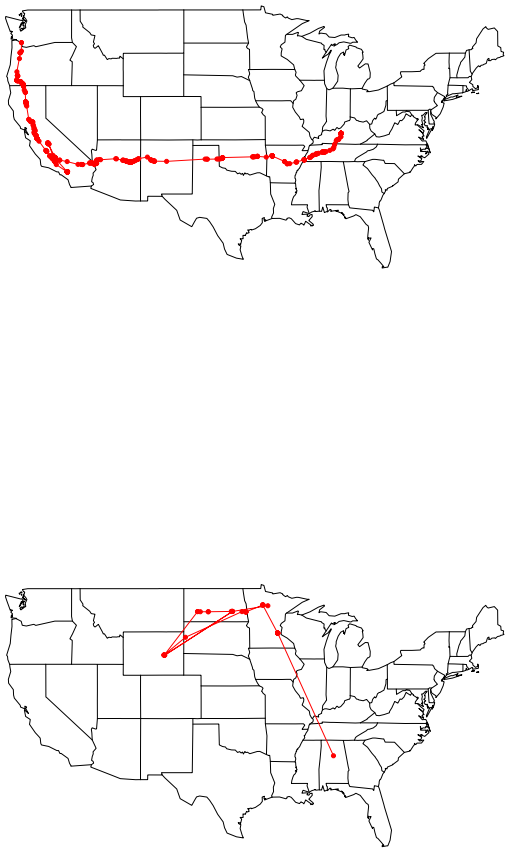

Figure 9: Positions of two users over the period of one week.

on geo-location, then more often than not you are close (within about 10 kilometers).

If we assume this is true even if the person mentioned does not have geo-location turned on, we can use this fact to obtain an estimate of the person's position. As we have noted, the best we can expect to do is on the order of 10 kilometers. This is adequate for many applications, such as detecting disease outbreaks, assessing the sentiment of a community toward some topic or event, detecting storms and power outages, etc.

As can be seen in Figure 9, one reason for the long tail in Figure 8 may be the movement of the individuals. We hypothesize that people who travel "tweet locally" in two different ways: they tweet back to their home, as if they haven't moved, but they also tweet to individuals who are close to where they currently are. We do not know how frequent this latter behavior is, but it seems logical that if one is visiting, one may tweet to people in the region visited, and this can help to detect people who are moving about.

Figure 10 shows what happens if we insist that mentions are both ways: we only compare positions between individuals who both mention the other. In this case, as can be seen by the inset, tweeters tend to tweet even more locally, although since the notches in the

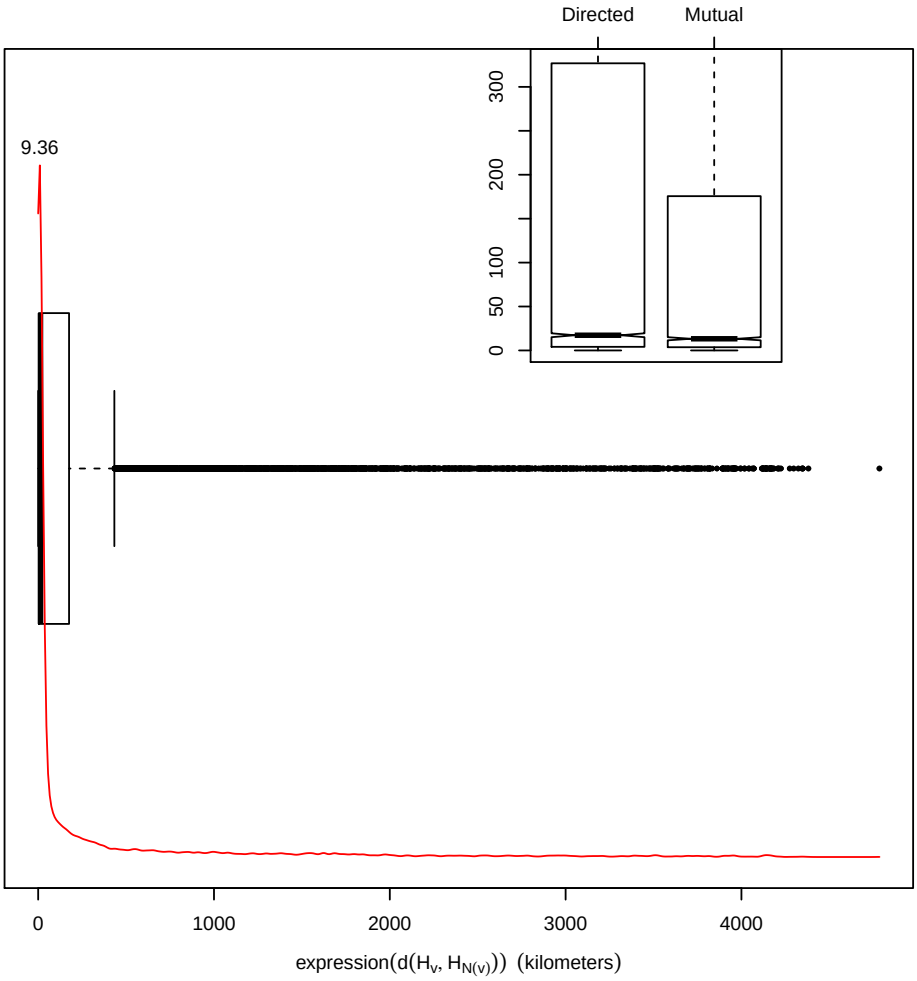

Figure 10: Distances between a user's positions and those of the people that are referenced by the user. In this plot, we only consider pairs in which each user mentioned the other – "mutual" mentions. These are the distances between home positions, as calculated using the 2 dimensional kernel estimator. The red plot shows the probability density of the distances (as estimated using a kernel estimator) with the distance associated with the maximum likelihood above the curve. There are 23298 distances represented in this plot. The inset shows the two boxes, from Figure 8 on the left and the mutual distances on the right, zoomed to the bulk of the data.

boxes overlap, the median distance is not significantly smaller. However, the bulk of the distances tend to be closer, by about half, when we insist on mutuality.

5. Discussion

This work is preliminary, and much needs to be done. In particular, fitting a model to data that is censured in this manner (we only observe those individuals who happen to geo-locate their tweets) is challenging. Fitting the random dot product model through the singular value decomposition has the problem that we are fitting the unobservables as if they are true non-edges. This biases the vectors to such a degree that they tend to be nearly 0 (the observed graph is so sparse that the low-rank approximations provided by the singular value decomposition are nearly 0).

It is clear that the geography can inform the inference. The simplest version would be to first group all the vertices according to their geography: use the home coordinates for each vertex and cluster these into groups. Then compute the group-to-group matrix to obtain the between group vectors. Within each group, compute the within-group vectors. Finally, combine these all into one model.

The above will only work for the subset of vertices for which we have locations, and so it would need to be modified to account for the unobserved vertices. This could be done by first inferring their coordinates, and grouping them, or by some ad hoc method of assigning the vertices to groups, such as the group they are most connected to.

We did not address the time series nature of these graphs. Preliminary indications are that there is a strong short-term correlation – if you tweet to Bill today you are very likely to tweet to him tomorrow. This makes sense since users have a collection of "real friends" that they communicate with regularly. This explains the local nature of tweets (your friends tend to be geographically close).

The geographic inference work is quite promising. Tracking users to within 10–100km is perfectly adequate for most applications such as public health monitoring, responses to products and advertising, or monitoring social unrest. The large volume of tweets make the analysis challenging, and the constraints placed on the analysis by the collection restrictions also provides interesting questions for further research.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, 2011.

Avanti Athreya, Vince Lyzinski, David Marchette, Carey Priebe, and Daniel Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya*, to appear, 2015.

Nigel Collier, Nguyen Truong Son, and Ngoc Mai Nguyen. OMG u got flu? analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*, 2:(Suppl 5):S9, 2011. URL `http://www.jbiomedsem.com/content/2/S5/S9`.

Courtney D. Corley, Diane J. Cook, Armin R. Mikler, and Karan P. Singh. Text and structural data mining of influenza mentions in web and social media. *Int. J. Environ. Res. Public Health*, 7, 2010. doi: 10.3390/ijerph7020596.

Mark Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27 (4):81–84, July/August 2012. doi: 10.1109/MIS.2012.76.

Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *CoRR*, abs/0812.1045, 2008. URL `http://arxiv.org/abs/0812.1045`.

Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector – tracking epidemics on twitter. *Machine Learning and Knoledge Discovery in Databases*, 6323:599–602, 2010.

Kalev H. Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), May 2013. doi: doi:10.5210/fm.v18i5.4366.

Kristina Lerman and Rumi Ghosh. Information contagion: n empirical study of the spread of news on digg and twitter social networks. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 90–97, 2010.

David Marchette and Carey Priebe. Predicting unobserved links in incompletely observed networks. *Computational Statistics and Data Analysis*, 52:1373–1386, 2008.