Assessing Connectivity Bias in the Twitter Network

Adam Bates and Lara Letaw
Department of Computer & Information Science
University of Oregon
{amb,zephron}@cs.uoregon.edu

Abstract—**Over the past few years Twitter has been at the forefront of the online social networking phenomenon. Twitter experienced a growth rate of almost 1400% between February 2008 and February 2009. This surge in popularity has legitimized Twitter as a channel for communities to interact within the United States and across the globe. Still in its nascent stages, there is much at that is unknown about the nature of interaction and influence within the Twitter network.**

This paper presents a detailed analysis of the nature of the relationships that exist between users within an online social network. We have collected a complete subgraph of the Twitter network that contains tens of millions of users and hundred of millions of those users' relationships. We profile this subgraph and explore its strengths and limitations by comparing it to a random sample of Twitter users. We then characterize the connectivity of our dataset based on the user attributes associated with each Twitter account. We use these results to identify a number of bias patterns that can be observed between different user groups. It is our belief that these observations shed new light on the nature of the global community's Twitter usage.

Keywords— **Online social network, Twitter, Measurement, Overlay topology**

I. Introduction

Twitter is a microblogging and social networking servive that allows its users to send short, one-to-many messages called *tweets*. Like many online social networks, users can opt-in to viewing other users' messages *following* them, creating a customized content stream. The act of following is an implicit endorsement of another user on the network. It can be inferred that the user being followed is of interest to another user on the network. These interactions constitute a networked system through which users connect and message propagate.

In order to gain a better understanding of Twitter, it is necessary to investigate the nature of this network. This information can be used when considering design decisions or allocating resources within Twitter or other future social networks. While Twitter is innovative in the manner in which it distributes messages, a great deal of this profiling must take place at the user-level network.

Twitter, like many other online social networks, present a novel opportunity to study interactions between people. Through them, it becomes possible to easily quantify the activities of millions. The simple and open structure of the Twitter network is particularly well-suited for this purpose. As each user account is associated with a variety of attributes, there a variety of methods to approach this task. The work we present here establishes a framework for further analysis.

There are many commonly held intuitions regarding influence that can be proved, disproved or measured through analysis of Twitter. For example, it is a commonly held belief that activity in states such as California and New York is of a higher culturally relevance than the activity in other states. As each user account is associated with a location field, it is possible to use the Twitter network to measure the influence of these allegedly trend-setting states. Other, subtler biases can also be identified and evalued in the same manner.

A defining aspect of Twitter that lends itself to research is that all activity defaults to public. Each user's messages are visible unless they explicitly change their privacy setting. Although tweets can be protected, all other information associated with an account visible to all. Using this publicly available connectivity information, it is possible to make broad assessments of connectivity within the Twitter community.

Given the massive size and continued growth of Twitter, capturing a complete snapshot of the network is an increasingly unrealistic endeavor. To make matters worse, Twitter imposes prohibitive rate limitations on access to many of its network measurement resources. It is therefore necessary to obtain a representative and meaningful sampling of the network before analysis begins.

One possible option is to take a random sample of a small percentage of Twitter accounts and activity. Inspection of this random sample's attributes would lead to a representative view, but not necessarily a meaningful one. The users of greatest interest are small in number and some of their attributes will have extreme values. A random sample is not likely to capture these users' impact on the network.

It is also possible to conduct a biased sampling. Here, interesting users with extreme user attribute values are specifically targetted for measurement. If good metrics are used to assess influence, this will ensure that important user accounts are not crowded out by unimportant ones. Of course, this leads to data that is less representative than a random sampling. It can be observed that there is a natural trade-off between these two goals.

Our approach finds a healthy compromise between these competing priorities by starting with a biased sample and then performing a mult-hop crawl across a small piece of the network to conduct further sampling. This crawl creates a snapshot of a part of Twitter that is acceptably representative while simultaneously ensuring that rare users are acceptably prominent. We establish that our snapshot is representative by comparing its profile to the profile of a true random sample.

STILL NEED TO PUT A SUMMARY OF OUR WORK

APPROACH AND MAIN FINDINGS HERE (AS PER REZA'S REPORT DESCRIPTION)

The rest of this paper is organized as follows. We outline our data collection process in section II. In section III, we present our method for evaluating connectivity bias within the Twitter network. Analysis of our results is contained in section IV. We review some of the related work in section V. Section VI concludes the paper.

II. Data Collection

At the start of our work we inherited a large dataset from Rejaie *et all* that was created through a biased sampling of the Twitter network. The dataset contained partial connectivity information for 242,275 potentially influential users. Due to limitations of the Twitter REST API at the time of collection, the connectivity data was incomplete. Additionally, the dataset did not include the associated attribute information for each user because it was not germaine to Rejaie's previous work. This dataset needed to be completed before beginning our analysis.

Our goal was to characterize the relationships in a completed section of the Twitter network. It was first necessary to determine if a large, complete subgraph existed within this dataset. A collection of smaller subgraphs would not be sufficient as they would be less likely to capture the impact of rare influential users. By performing a series of reconstructions of the graphs in the dataset we were able to discover a subgraph of 215,606 sampled users. Including the next hop information for these sampled users, the subgraph included 15,548,091 unique user accounts.

The inherited dataset was not originaly intended to measure the extreme connectivity degrees of rare users. As a result, the friend and follower counts for each user in the subgraph were truncated at 1,000. This truncation was due to the prohibitive rate limiting of of the Twitter API, which made it infeasible to collect the complete connectivity information for these users in real time. Over the course of several weeks, we were able to restore this missing information. Several months passed between these two collection phases. The potential error that this introduced is explored below.

Simultaneously, we needed to collect the user attribute information for the tens of millions of users in our subgraph. Completing this task via the Twitter API was difficult due to the connectivity restoration process, and it was not our wish to violate Twitter's terms of service by launching API calls from dozens of hosts. Fortunately, we were able to find an alternate method of collecting user attribute information. Twitter provides XML dumps of user information through their primary website. These calls are not subject to API rate limites, presumably for the benefit of mobile applications that may issue frequent user lookups. Using this service, we were able to collect the necessary information without violating the Twitter API ToS.

A. Measuring Error

Because our user attribute information was collected in less than a month, we have assumed that no significant error is present. It is not possible to capture a realtime snapshot of user attributes in the Twitter network, so we have no method of comparing collected attributes to actual attributes. Our intuition here is that most user attributes are not subject to frequent change. The exception to this is friend and follower count, which is explored below. For our purposes, the collected user attribute information can be accepted as accurate.

In contrast, connectivity information for our subgraph was collected in two different phases between September 2010 and February 2011. It would be foolish to assume that significant change did not occur in the network over this time. It was therefore necessary to confirm that the temporally disparate collection phases did not introduce error to our network measurements. To do this, we compared our snapshot of connectivity for each user to the stated friend and follower counts in each user's attributes.

GRAPHIC for FOLLOWERS: The CDF I am working on that shows the error percentage for our core users between actual degree and measured degree.

GRAPHIC for FRIENDS: The CDF I am working on that shows the error percentage for our core users between actual degree and measured degree.

Figures **??** and **??** confirm that we have captured an accurate view of connectivity in our piece of the Twitter network. 90% of users blah blah blah. This speaks to the general stability of the social network that Twitter facilitates.

(Q: Plotting the growth of this subgraph – would this be a neat consideration for future work? Not that we need to do it, but we can always suggest it in the paper for the benefit of Team Reza)

B. Profiling

Paragraph: Restate what we said in the introduction about how we need to compare our subgraph to a random subgraph in that magical way that papers manage to do.

GRAPHICS FOR EACH ATTRIBUTE!

Paragraphs summarizing the attributes.

Paragraph: Conclude that our dataset is sufficiently representative of a random twitter sampling to warrant further investigation. Make note of any attributes for which this is not the case.

III. Methodology

A. Approaching Attribute Analysis

Although each user account is associated with 33 publicly available attributes, only a few of these were of any interest for our purposes. 10 of the attributes are profile display settings, which were not considered. Another 6 are user identified such as name and ID which are not of any use to network-level analysis. We initially identified 14 potential attributes of interest. Further description of these attributes is available in appendix VI.

After selecting attributes for consideration, we identified appropriate groupings for each attribute. The format of the considered attributes took one of three forms. These formats informed the manner in which they were grouped. Many of the attributes had a finite number of options, such as a boolean value for *Protected* or one of seven options for *Language*. A group was assigned for each possible value of these attributes.

The next field format was that of a continuous integer range, such as *Follower Count* or *Status Count*. For these, groups took the form of a range of values. It was important to create groups in such a manner that rare, high values user were clearly visible in the results. It was also important that the large population of low value users were represented. To accomplished this, we increased the range of our groupings logarithmically.

The final and most challenging format took the form a free text field. Free text fields are used by the *Bio*, *URL* and *Location* attributes. Of these, only *Location* was considered in our study. After parsing this field for identifiable locations on a variety of scales, it was decided that groupings by U.S. state would be the most impactful. Of these, results for the top ten most popular states are included in this paper.

Paragraph: Mention the simple random survey of records that was performed, the percentage of users with no entry, the percentage of users with an indecipherable entry, etc.

After identifying discrete groupings for each attribute it became possible to identify attribute relationships within our dataset. Each edge in our subgraph represents a relationship between two Twitter users, each with an group assignment for each attribute. We created a unique identifier for all possible relationships for each attribute. We then processed our dataset by assigning these identifiers and counting their frequency.

B. Measuring Bias

Through data processing we obtained a set of values A_{XY}, representing the actual count of each possible attribute relationship *X follows Y*. However, this value alone is insufficient for measuring network bias. It does not consider the frequency of users of attribute X or Y. It also does not consider the total number of outgoing edges from X, X_{out} ,or incoming edges to Y, Y_{in}. Before network bias can be determined it is necessary to normalize A_{XY} against the value one would expect in a randomized, bias-free version of the graph. This value is given by:

$$R_{XY} = \frac{X_{out} * Y_{in}}{\sum_{x,y,z} u}$$

where X_{out} is the number of outgoing edges from X, Y_{out} is the number of incoming edges to Y, and $\sum_{x,y,z} u$ is the sum of all edges for all attribute values in the network. For our Twitter network, R_{XY} is consistent when considered from both the incoming and outgoing edge perspective because of following fact:

$$\sum_{x,y,z} u_{out} = \sum_{x,y,z} u_{in}$$

Given A_{XY} and R_{XY}, the bias for the relationship *X follows Y* is given by:

$$B_{XY} = \frac{A_{XY} - R_{XY}}{R_{XY}}$$

A positive B_{XY} indicates that group X is biased towards following group Y. A negative B_{XY} indicates that group X is biased against following group Y. If $[B_{XY}]$ is very nearly zero, no bias is observed. A $[B_{XY}] > 1$ indicates that the relationship *X follows Y* is biased to the point that it is several times more frequent or infrequent than would be expected in a randomized graph.

IV. Analysis

Paragraph – introduce results.

A. System Age

B. State

C. UTC Offset

D. Language

E. Geo-Enabled

F. Protected

G. Follower Count

H. Friend Count

I. Statuses Count

J. Listed Count

Include commentary on what the graphs tell us throughout.

Conclude that across these attributes there are several identifiable patterns – similar users follow eachother, dissimilar users follow eachother, low degree users follow high degree users, extreme low degree users deviate from general usage, extreme high degree users deviate from general usage. Identify which attributes belong to which patterns.

V. Related Work

One of Twitter's core functions is the analysis of individual messages to determine trending topics. As such, Twitter is capable of acting a cultural barometer. A variety of studies have attempted to answer the question of how information spreads within the Twitter network. Investigations have included spam, the meme life cycle, and news discovery and explanation.

A variety of Twitter-related work has been conducted over the past several years. The majority of this work focuses on the analysis and propagation of individual messages as they spread through the network. Cheng *et al* have used tweets to attempt to geo-locate users based on the content of their messages [1]. Lerman and Ghosh have studied Digg and Twitter to measure news items' lifespan and speed within social networks [2]. Sadikov and Martinez have conducted similar work, chosing instead to focus on URL and tag propagation [3].

It is only reasonable that message-level analysis has captured the attention of the research community. After all, the novelty of modern social networks is largely that they are a new medium for the spread of information. For the purposes of assessing influence, however, message-level analysis does not tell the whole story. Much of this work lies in predictive message filtering or spotting trending phrases. This is not perfectly suited for the task of broadly determining who is influencing who. Instead, we ask a much simpler question – what type of people are users choosing to listen to?

For this, we turn to user-level analysis. When a user choses to opt-in to another user's tweet stream, this says much more about influence than the propagation of individual messages. This idea of influence through followers is a truth that rests at the very core of Twitter, one that can be plainly seen by visiting any user's page and making note of the prominently displayed "Follower" and "Following" numbers. Our contribution will be to take this fundemental concept and cross-reference it with various user attributes in an attempt to make more general claims about the nature of influence in the Twitter network.

VI. Conclusion

In this paper we have presented a detailed analysis of a large subgraph of the Twitter network. We have identified a number of influence patterns across different attributes within the twitter network. Our work is unique in that it provides a method of tracking information flow without being dependant on access to individual messages. There are a number of pitfalls to message analysis that our work avoids. It is our belief that a combination of both methods of Twitter network analysis are necessary in order to gain a complete perspective on the way this tool is used.

Future work – ideas to consider regarding future projects that could come of our work?

APPENDIX

User Attributes Below are the user attributes we plan to examine in our analyses. In reference to number of baskets, *varied* means we will experiment with different numbers of baskets.

Attribute	**Description**	**# of Baskets (Core)**	**# of Baskets (All)**	**Values**
Location	User-reported geographic location.	Varied	Varied	(none), city, state/province, country, continent
Protected	If true, only approved followers may see the user's tweets	2	2	true, false
Followers Count	Number of users who track this user's tweets	Varied	Varied	integer
Friends Count	Number of users this user follows	Varied	Varied	integer
Account Creation Date	When this user first entered the system	Varied	Varied	date
Geo-Enabled	GPS meta data is included on tweets	2	2	true, false
Verified	Twitter has verified the identity of the user, currently used for Twitter partners and advertisers	2	2	true, false
Statuses Count	Number of tweets	Varied	Varied	integer
Language	User's chosen language	7	7	en, de, it, es, ja, fr, ko
Contributors Enabled	If enabled, multiple users can tweet from this account	2	2	true, false
Listed Count	Number of lists that include this account	Varied	Varied	integer
Show All Inline Media	Display photos and videos of other users, not just friends	2	2	true, false
URL	This user posted a URL	2	2	true, false
Is Translator	User has signed up to translate other people's tweets	2	2	true, false

I. DATASET PROFILE

Available users are those for which user profile data was successfully collected. Unavailable users are those for whom an attempt to acquire user attributes resulted in a *404 Not Found* error. We assume these user accounts are closed.

	Total	**Available**		**Unavailable**	
Data Set	#	#	%	#	%
All	17,688,493	17,475,570	98.8	212,923	1.2
Core	242,275	238,323	98.8	3,952	1.2
Connected	15,548,091	14,624,837	94.1	923,254	5.9
Random	1,805,758	1,529,897	84.7	275,861	15.3

A. Subgraphs

Graph edges refers to the connections between core users and any other user.

Users in Subgraph 1	15,548,091
Available Users in Subgraph 1	15,363,277
Core Users in Subgraph 1	194,004
Remaining Core Users	48,267
Remaining Subgraphs	46,403
Connections to Subgraph 1	46,924
Total Connected Users	240,932

B. *Boolean User Attributes*

Number and percentages of users with a *true* value for boolean attributes.

```
SELECT COUNT(*) FROM table WHERE attribute=’true’;
SELECT COUNT(*) FROM table WHERE url!=’’ and url is not NULL;
```

	Core		**Connected**		**Random**	
Attribute	#	%	#	%	#	%
Protected	8,517	3.6	1,731,662	11.8	126,261	8.3
Geo-Enabled	58,432	24.5	2,573,007	17.6	101,937	6.7
Verified	119	0.1	933	0.0	67	0.0
Contributors Enabled	7	0.0	115	0.0	7	0.0
Show All Inline Media	18,870	7.9	788,201	5.4	25,080	1.6
URL	118,364	49.7	5,214,015	35.7	156,512	10.2
Is Translator	72	0.0	1,479	0.0	17	0.0

C. *Enumerated User Attributes*

Attributes for which there are a relatively small number of values, other than boolean values.

C.1 Language

Percentage of users per language.

```
SELECT lang, COUNT(id) FROM table GROUP BY lang;
```

		Core		**Connected**		**Random**	
Code	**Language**	#	%	#	%	#	%
en	English	173,782	72.9	11,553,068	79.0	1,272,730	83.2
es	Spanish	16,210	6.8	1,426,063	9.8	147,739	9.7
ja	Japanese	44,499	18.7	1,360,341	9.3	73,026	4.8
fr	French	1,222	0.5	111,423	0.8	17,665	1.2
de	German	1,560	0.7	107,880	0.7	10,923	0.7
it	Italian	713	0.3	54,948	0.4	7,178	0.5
ko	Korean	337	0.1	11,114	0.1	634	0.0

D. Location

The location attribute is a free-form value. That is, a user can set their location to anything they want. In order to the use the location data, we must filter and interpret these values. The goal is to map a single geographical location to each user, but the location data is not well-formed. Location data is sometimes not entered, non-specific (Ex: “Under your bed"), multiple (Ex: “NYC & San Francisco"), or simply a non-loaction (Ex: “Blahhhhh").

Our approach is to match each user’s location against a list of known valid locations. We are getting the list of locations from the World Cities Database (`http://www.maxmind.com/app/worldcities`). World Cities has a list of all cities, regions, and countries in the world. We first tried to match the beginning of each user’s location attribute with a city. That is, the user location could match the World Cities location exactly, or it could include characters beyond the World Cities location. Also, capitalization is ignored. We performed similar searches with region and country. Overall, there were approximately 7 million users matched with a location, out of the 14.5 million total.

Some users were matched with more than one location. For these users, we need to make a more specific location match by appending to the location string. For example, if a user matches many cities with the name “Springfield", we can search for “Springfield, OR", or “Springfield, Oregon", or “Springfield, IL", etc.

For the users that do not match a location, we need to examine why. We need to estimate what proportion of the non-match are non-locations, non-entered, or ill-formed. This will be done by manual examination of a subset of the data. We need to determine an accectable size for this subset.

Match Criterion	**Number of Users**
One or more cities	8,553,747
Exactly one city	983,516
One or more regions	2,442,249
Exactly one region	1,771,560
One or more countries	1,175,514
Exactly one country	1,164,406 (99.1%)

D.1 Other Approaches

The Google Maps and Yahoo Maps Geocoding APIs provide translation from user-entered location strings to geographical locations. For example, Google Maps knows that “The City by the Bay" is San Francisco. There are obvious advantages of this functionality, but we are unable to use these services because of strict rate-limiting. The Geocoding API is rate-limited to 2,500 queries per day per IP address, and the Yahoo Geocoding API has a 5,000 queries per day limit.

References

[1] Z. Cheng, J. Caverlee, and K. Lee, "You are whare you tweet: A content-based approach to geo-locating twitter users," in *Proceedings of the CIKM*, 2010.

[2] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks," in *Proceedings of the AAAI Conference on Weblogs and Social Media*, 2010.

[3] E. Sadikov and M. Martinez, "Information propagation on twitter," in *Proceedings of the ACM*, 2009.