CIS 632 - Technical Report: Twitter Project

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

I. Motivation

Twitter is a not just a new technology; it is also a new form of communication. Although there is a variety of ongoing Twitter-related research, there are a myriad of remaining potential approaches. There are still open questions regarding influence within the Twitter network. We hope to evaluate influence as it relates to the various attributes associated with a user account.

II. 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.

III. Methodology

We have acquired a dataset of potentially influential twitter accounts from the previous work of Rejaie *et al*. This assessment was based on connectivity, user age, and other factors. After building our subgraph and collecting the related user attribute information, we intend to filter the subgraph based on these attributes. We hope to be able to report on the interesting interactions between the different sets of attribute-sorted users.

Our goal will be to evaluate the nature of the connectivity of our Twitter subgraph. We intend to discover connectivity biases amongst users of different attribute groups. It is our belief that there is a relationship between a given user's attributes and the likelihood of that user being connected to other users of various attributes. For a simple example, we predict that there will be a relationship between a user's location being Dallas, TX and the likelihood that other connected users are also from Dallas, TX.

To assess connectivity bias, we will group each user based on a given attribute into baskets of similar users. There are four different forms of connectivity that must be considered – intra-outgoing (followers within basket), intra-incoming (friends within basket), inter-outgoing (followers outside of basket) and inter-incoming (friends outside of basket). The two intra-connectivity measurements will be assessed for each basket. The two inter-connectivity measurerments will be assesed from each basket to every other basket.

We will then compare these results to internal and external randomly selected baskets of users. The internal comparison will be against our main dataset. This comparison will tell us how significant the connectivity bias for a given attribute is within our tiny Twitter universe. The external comparison will be against another set of users that were selected randomly without regard for their influence. This will tell us how significant the measured connectivity bias is relative to Twitter as a whole. Using two different random distributions will also help to validate the relevance of our results to the entire Twitter network.

The above process will be repeated for as many user attributes as is prudent for the scope of our project.

A. Creating Subgraphs

After inheriting our dataset, the first step was to determine the size and nature of the Twitter subgraph that we were inspecting. To do this, we ran a script that crawled across the entire dataset and assigned each entry a subgraph ID. We discovered that the vast majority of dataset, 194,004 entries, were already connected. It logically followed that the vast majority of the 17,688,493 unique user IDs in our dataset were also already contained in the primary subgraph. As Twitter boasted 190 million users during the summer of 2010, 17.5 million represents a non-negligible amount of Twitter's active users.

B. Shortcomings of Inherited Dataset

Due to the stringent rate limits imposed by the Twitter REST API, connectivity information for our inherited dataset was truncated at 1,000 friend IDs and 1,000 follower IDs per user. This led to our data exhibiting two unfortunate traits. First, because connectivity data was incomplete, inconsistencies would arise in which one user's entry did not reflect the connection that was claimed by another user's entry. Second, we were only looking at a fraction of the picture for our high degree users. Presumably, these users would be some of the most relevant and valuable members of our graph. After discovering that 28,669 of our user entries had been truncated at 1,000 for either their friends or followers, we took the following steps to improve our dataset.

First, we attempted to bridge together the subgraphs of our dataset. When one user entry did not reciprocate the claimed connection of another user from a different subgraph, it was due to the connectivity truncation described above. Whenever this occurred, we made a note of the locations at which our subgraphs were being bridged together. We then went back and augmented the entries in the primary subgraph to include the users on the other side of the bridges. In doing so, we were able to increase the number of users in our primary subgraph from 194,004 to XXX,XXX.

Second, we went back to the Twitter REST API to reclaim our missing connections. We were able to obtain complete connectivity data for XX,XXX of our 28,669 truncated core users. The remaining accounts had been disabled, suspended, or were otherwise unreachable through Twitter. The main dataset was then augmented with the new connectivity data for our high degree users. In addition to giving us a more accurate view of our core user's connectivity, this added X,XXX,XXX edges to our primary subgraph.

While both of these steps increased the level of information in our data, we recognize the potential inconsistencies that these measures introduced. The most obvious is the fact that our high degree user's friends and followers were very likely to have changed between late 2010 and early 2011. Additionally, selectively bridging the subgraphs within our dataset could potentially introduce bias in our connectivity analysis.

Despite the patchwork nature of our dataset, we feel that our findings are plausibly representative due to the nature of our analysis. This is mostly due to the fact that we investigating connections within a subgraph, not the user nodes themselves. For this purpose, we believe that amassing as many graph edges as possible will deliver the most meaningful results. Additionally, the anachronisms introduced by merging multiple collection phases are only a minor concern because we do not plan to plot changes in influence over time. There is a certain level of timelessness to the connections that are reflected in our dataset. For example, if a follower of Ashton Kutcher's subsuently chose to stop following him, there is still information to be gleaned from that past connection. For this reason, we were comfortable proceeding forward in spite of inconsistencies in a small minority of our data.

IV. Evaluation

We will our connectivity dataset and crawl across it to find a particularly great subgraph. If time allows, we will attempt to restore and expand the incomplete information within this subgraph. We will then create overlays of this subgraph using different user attributes. We will analyze the directional nature of these subgraph overlays.

V. Timetable

(To be revised as needed.)

January 3-23

- Developed project proposal
- Determined feasibility of data collection
- Built subgraphs of core users
- Collected user data for all core and leaf users

January 24-30

- Ensure data set is complete as far as user attribute collection
- Perform preliminary user attribute distribution analysis
- Bridge subgraph connections
- Determine feasibility of collecting missing friends and followers
- Acquire random data sets
- Build baskets for which baskets are already enumerated

January 31 - February 5

- Formalize method for characterizing connectivity
- Determine basket filters for variable attributes
- Determine basket sizes for continuous attributes
- Merge subgraph and user attribute databases

February 6-12

- Select most informative basket configurations
- Select most interesting user attributes
- Determine if there are any interesting composite attributes to inspect
- Begin creating connectivity visualizations

February 13 - March 1

- Continue to refine and polish attribute basketing and assessment
- Complete first draft of research paper

March 2-13

- Complete final draft of research paper

March 14

- Submit final research paper
- Give project presentation

VI. 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
UTC Offset	Time offset from Coordinated Universal Time	34	34	integer
Time Zone	Logitudinal region	141	143	time zones
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
Status Source	Where the most recent status was tweeted from	5,144	39,566	(none), web, various URLs

VII. 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 Users	17,688,493
Total Core Users	242,275
Total Available Users	13,877,912
Total Available Core Users	238,323 (98.8% of core)
Total Unavailable Core Users	3,952 (1.2% of core)

A. Subgraphs

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

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

Attributes whose values are either *true* or *false*.

Attribute	# True (Core)	% True (Core)
Protected	8,517	3.575
Geo-Enabled	58,432	24.518
Verified	119	0.050
Contributors Enabled	7	0.003
Show All Inline Media	18,870	7.918
URL	118,364	49.665
Is Translator	72	0.030

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.

Language Code	Language	# (Core)	% (Core)
en	English	173,782	72.919
ja	Japanese	44,499	18.672
es	Spanish	16,210	6.802
de	German	1,560	0.655
fr	French	1,222	0.513
it	Italian	713	0.299
ko	Korean	337	0.141

C.2 UTC Offset

UTC Offset	# (Core)	% (Core)
(none)	38287	16.065
32400	36551	15.337
-18000	26196	10.992
-10800	25001	10.490
-28800	23344	9.795
-21600	17830	7.481
-14400	12698	5.328
25200	10425	4.374
-36000	10162	4.264
3600	9607	4.031
0	7067	2.965
-25200	4360	1.829
28800	4065	1.706
-32400	3573	1.499
-16200	2685	1.127
7200	1908	0.801
36000	1286	0.540
10800	1054	0.442
19800	717	0.301
43200	322	0.135
-39600	236	0.099
14400	213	0.089
18000	163	0.068
12600	138	0.058
34200	103	0.043
-7200	100	0.042
21600	91	0.038
-12600	57	0.024
46800	24	0.010
-3600	20	0.008
39600	19	0.008
16200	9	0.004
23400	7	0.003
20700	5	0.002

C.3 Time Zone

Only the top 100 time zones are shown here.

Time Zone	**# (Core)**	**% (Core)**	**Time Zone**	**# (Core)**	**% (Core)**
(none)	38287	16.065	Kyiv	215	0.090
Tokyo	30332	12.727	Brussels	214	0.090
Pacific Time (US & Canada)	23280	9.768	Monterrey	214	0.090
Brasilia	15997	6.712	Lima	194	0.081
Central Time (US & Canada)	15327	6.431	Copenhagen	183	0.077
Eastern Time (US & Canada)	15066	6.322	Riyadh	178	0.075
Santiago	12131	5.090	Abu Dhabi	173	0.073
Hawaii	10162	4.264	Athens	164	0.069
Quito	10152	4.260	Bucharest	151	0.063
Jakarta	9334	3.917	West Central Africa	151	0.063
Greenland	7968	3.343	Warsaw	148	0.062
London	5968	2.504	Perth	145	0.061
Mountain Time (US & Canada)	3826	1.605	Chennai	144	0.060
Amsterdam	3600	1.511	Auckland	139	0.058
Alaska	3573	1.499	Tehran	138	0.058
Osaka	3509	1.472	Guadalajara	133	0.056
Caracas	2685	1.127	Cairo	130	0.055
Seoul	1912	0.802	Vienna	126	0.053
Mexico City	1829	0.767	Wellington	124	0.052
Singapore	1677	0.704	Kuwait	122	0.051
Berlin	1494	0.627	Jerusalem	117	0.049
Madrid	1160	0.487	Budapest	115	0.048
Paris	1106	0.464	Bern	114	0.048
Buenos Aires	1008	0.423	Mid-Atlantic	100	0.042
Bangkok	1000	0.420	Helsinki	96	0.040
Kuala Lumpur	944	0.396	Riga	93	0.039
Sapporo	762	0.320	Adelaide	91	0.038
Hong Kong	627	0.263	Hanoi	73	0.031
Moscow	595	0.250	St. Petersburg	73	0.031
Rome	594	0.249	Belgrade	71	0.030
Sydney	562	0.236	Nairobi	66	0.028
Bogota	526	0.221	Tijuana	64	0.027
Istanbul	474	0.199	Casablanca	63	0.026
Arizona	455	0.191	Ekaterinburg	63	0.026
Edinburgh	411	0.172	Prague	62	0.026
Beijing	399	0.167	Newfoundland	57	0.024
Melbourne	396	0.166	Minsk	54	0.023
Dublin	362	0.152	Sofia	51	0.021
Stockholm	357	0.150	Islamabad	49	0.021
Atlantic Time (Canada)	339	0.142	Kolkata	49	0.021
New Delhi	304	0.128	Canberra	48	0.020
Central America	301	0.126	Fiji	46	0.019
Pretoria	275	0.115	Chihuahua	44	0.018
Indiana (East)	258	0.108	Harare	44	0.018
Lisbon	243	0.102	Karachi	43	0.018
Taipei	231	0.097	Zagreb	39	0.016
La Paz	228	0.096	Yakutsk	36	0.015
Brisbane	225	0.094	Mazatlan	35	0.015
Mumbai	220	0.092	Novosibirsk	32	0.013
International Date Line West	216	0.091	Georgetown	28	0.012

C.4 Status Source

Source of the most recent tweet. Only the top 50 sources are included here.

Source URL	Source Name	# (Core)	% (Core)
(none)	web	65604	27.527
www.ubertwitter.com/bb/download.php	ÜberTwitter	23752	9.966
twitter.com/	Twitter for iPhone	11737	4.925
blackberry.com/twitter	Twitter for BlackBerry®	11082	4.650
www.tweetdeck.com	TweetDeck	10571	4.436
(none)	(none)	9726	4.081
twitterfeed.com	twitterfeed	8734	3.665
www.echofon.com/	Echofon	6847	2.873
mobile.twitter.com	Mobile Web	6613	2.775
twitter.com/devices	txt	4681	1.964
twtr.jp	Keitai Web	4226	1.773
twittbot.net/	twittbot.net	3375	1.416
z.twipple.jp/	ついっぷる/twipple	3270	1.372
www.movatwi.jp	www.movatwi.jp	3006	1.261
www.hootsuite.com	HootSuite	2527	1.060
mobile.twitter.com	Twitter for Android	2520	1.057
www.snaptu.com	Snaptu	2323	0.975
www.tumblr.com/	Tumblr	2201	0.924
www.facebook.com/twitter	Facebook	2064	0.866
twidroyd.com	twidroyd	2060	0.864
www.google.com/support/youtube/bin/answer.py?hl=e	Google	1974	0.828
sourceforge.jp/projects/tween/wiki/FrontPage	Tween	1712	0.718
twitter.com/tweetbutton	Tweet Button	1698	0.712
m.tweete.net	m.tweete.net	1677	0.704
tinyurl.com/tweetcaster	TweetCaster	1574	0.660
foursquare.com	foursquare	1424	0.598
www.nibirutech.com	TwitBird	1303	0.547
twicca.r246.jp/	twicca	1193	0.501
yubitter.com/	yubitter	1189	0.499
www.flight.co.jp/iPhone/TweetMe/	TweetMe for iPhone	904	0.379
www.twittascope.com	Twittascope	868	0.364
levelupstudio.com	Plume	814	0.342
dlvr.it	dlvr.it	805	0.338
jigtwi.jp/?p=1	jigtwi	791	0.332
twipple.jp/	ついっぷる for iPhone	696	0.292
itunes.apple.com/us/app/twitter/id409789998?mt=12	Twitter for Mac	685	0.287
projects.playwell.jp/go/Saezuri	Saezuri	674	0.283
itunes.apple.com/app/twitter/id333903271?mt=8	Twitter for iPad	591	0.248
www.osfoora.com	Osfoora for iPhone	574	0.241
formspring.me	Formspring.me	571	0.240
stone.com/Twittelator	Twittelator	561	0.235
m.tuitwit.com	Tuitwit	548	0.230
twitpic.com	Twitpic	525	0.220
m.dabr.co.uk	Dabr	493	0.207
www.movatwi.jp	モバツイ	479	0.201
twitterrific.com	Twitterrific	477	0.200
(none)	Keitai Mail	463	0.194
www.socialoomph.com	SocialOomph	436	0.183
twtkr.com	twtkr	423	0.177
www.myspace.com/sync	MySpace	422	0.177

D. Location

Location data is entered free-form. That is, a user can set their location field to anything they want. In order to use the location data, we need to figure out which locations are not valid, and we need to transform valid locations into a format that can be processed. For example, we want all users from San Francisco to have a location attribute of *San Francisco, CA, USA*, so that we can easily determine which users share a location. We are using the Google Maps Geocoding API to make this conversion. When the API returns more than one result, we can either remove those location from our analysis, try a different method of processing, or try to determine the commonality between multiple results (if, for example, all results are in France, we can use France in our analysis). When the API returns zero results, we cannot use the location data.

The Geocoding API has some amount of error. The returned results are not always a correct transformation of the original location. We cannot manually examine each result for error, but we do plan to look at a subset of the results and derive a general error rate from that sample.

The Geocoding API is rate-limited to 2,500 queries per day per IP address. Google Maps API Premier members can query up to 100,000 per day, but that service starts at $10,000, so is probably beyond the budget of this project. Yahoo also has a Geocoding API, to which we can make 5,000 calls per day. However, we would prefer not to complicate the data collection by using both Google and Yahoo services.

Rate-limiting of Geocoding queries is an issue we need to address urgently. We either need to find a way to make more requests to the API without breaking the terms of service, or we need to get the information from the Google or Yahoo front-end. Alternatively, we could design our own location-conversion script, but we see this as non-ideal.

REFERENCES

[1] Z. Cheng, J. Caverlee, K. Lee, *You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users*, CIKM'10, October 26-30, 2010. Toronto, Ontario, Canada.
[2] K. Lerman, R. Ghosh, *Information Cntagion: An Empiral Study of the Spread of News on Digg and Twitter Social Networks*, AAAI Conference on Weblogs and Social Media 2010, 90–97.
[3] E. Sadikov, M. M. M. Martinez, *Information Propagation on Twitter*, ACM 200X.