HOMEWORK ROUTE FORM

Stanford Center for Professional Development Student Information

(Course No.): CSE 224W
(Faculty / Instructor Name): Jure Leskovec
(Date): 12/11/2011
(Student Name): Peter Lipay
(Phone): 206-708-9581
(Company): Microsoft
(Email): plipay1@gmail.com
(City.): Mercer Island
(State): WA
(Country): United States

Check One:
Final Project

The email address provided on this form will be used to return homework, exams, and other documents and correspondence that require routing.
(Total number of pages faxed including cover sheet): 13

For Stanford Use Only
Date Received by the Stanford Center
For Professional Development:

Date the Stanford Center for
Professional Development returned
graded project:

Date Instructor returned graded project:

S T A N F O R D U S E O N L Y

Score/Grade:

(to be completed by instructor or by teaching assistant)

Please attach this route form to ALL MATERIALS and submit ALL to:
Stanford Center for Professional Development
496 Lomita Mall, Durand Building, Rm 410, Stanford, CA 94305-4036
Office 650.725.3015 | Fax 650.736.1266 or 650.725.4138
For homework confirmation, email scpd-distribution@lists.stanford.edu
http://scpd.stanford.edu

Last modified October 27, 2008

Investigating User Communities on Twitter

Peter Lipay (plipay@stanford.edu)

Abstract: Many researchers have proposed automated ways of clustering users based on their relationships, specifically in the digital space. Likewise, there have been numerous data extraction techniques and metrics to pull significance and meaning out of vast amounts of user text. However, the two approaches have never been combined in a giant digital space like the Twitter Social network. In this study, I apply a Louvain clustering algorithm on a graph built up from a subset of the Twitter Social graph to partition Twitter users into distinct communities. I then apply automatic data-labeling techniques to these communities and attempt to generate meaningful labels for them. I also check to see if self-reported geographic location is a strong factor in bringing these communities of users together. This initial trial-run of my combined approach appears to be both possible and successful, providing useful and interesting data labels for the communities. I also find that geographic location does not appear to be a very strong factor in tying communities together.

1. Introduction

The Twitter network has approximately 100,000,000 active users, according to their CEO Dick Costolo [1]. Users post short status messages, called "Tweets", on their publicly displayed webpage for the entire world to see. Users can interact by choosing to "follow" one another, allowing them to easily see the updates of other users, in addition to various more direct communication methods. This mass of users forms a gigantic interconnected social graph, with a long series of factors determining how individuals connect and interact with one another. What also inevitably emerges in such a social graph is a set of distinct sub-communities within the overall social graph, driven by these user interactions. While existing studies have attempted to cluster random user samples from Twitter into distinct communities, I aim to go one step further and attempt to investigate the characteristics of these communities, and hopefully find some of the common interests or attributes that bind these groups together. Such an approach, if successful, would have wide-ranging potential in both academic and commercial spheres, from determining what interests drive social groups in society, to developing more sophisticated ad-targeting techniques.

2. Related Work

One of the most recent studies exploring clustering algorithms on Twitter is a study by Pujol et al. [2] from 2009. In it, they evaluate the effectiveness of several partitioning algorithms, based off of the Louvain method for detecting communities. The principle tactic is to partition the social graph in a way which maximizes the "modularity", which they define as "the difference between the number of edges within communities and the expected number of such edges", and they evaluate these algorithms on the Twitter and Orkut social graphs. However, while they do explore efficient ways of identifying communities within these networks, the aim is purely focused around optimizing website performance on these systems by putting different communities onto different servers, and no real investigation of the characteristics of these communities is performed.

A recent informal study that did look into what kinds of Twitter users are out there was done by Kalafatis [3] in 2009. In it, he harvested the biographies of a large dataset of Twitter users and subdivided them into communities by looking for occurrences of several pre-defined keywords, such as "Geek", "Parent", and "Business Owner". Then, within these communities, he came up with a list of associated terms that correlated strongly for each community group. However, this study starts out by arbitrarily choosing community labels to divide users into these clusters, while I aim to do the exact opposite. To identify clusters of users using the structure of the Twitter social graph itself, and then automatically generate labels for these clusters using data from those user's biographies and posting histories.

3. Methodology

My study uses a combination of Louvain community detection and word frequency analysis to automatically detect and label communities within the Twitter social graph. I have also cross-referenced this with the self-reported geographic locations of users in these communities to see if location is a significant factor tying communities together. The entire process consists of three distinct phases:
Data Collection, Community Identification, and Group Labeling.

Data Collection:

This study has been carried out through extensive use of the Twitter REST and Stream APIs [4], which allow easy developer access to information about users inside the Twitter social graph. I perform my analysis using a graph built up using data collected from the Twitter social network, aiming to model the connections of its users. Graph building and analysis is performed using python with the NetworkX library [5]. I represent a given Twitter user as a node in my graph, and connect two nodes with an undirected edge if I consider them to be "friends".

The first challenge here is how to define two users as being "friends". While users often have upwards of a hundred or more followers, as Huberman et al. [6] show, a given user only ever directly communicates (by mentioning their Twitter username with the "@" symbol) with about 10% of the people they're following. As I am interested in trying to map clusters of related users with presumably similar interests and attributes, I want edges to represent a somewhat significant connection between the two users.

The initial definition I planned to use was the following:

Given a user A and a user B, I propose to define users A and B as friends if and only if:

user A and user B are "mutual follows" (user A is following user B and user B is following user A) , AND
user A has "mentioned" (by referencing their Twitter id using the "@" symbol) or retweeted
user B at least twice, AND
user B has mentioned or retweeted user A at least twice.

However, after applying this to some initial test data, I quickly realized that these conditions were simply far too restrictive. Due to rate-limiting on the Twitter APIs, I limit myself to only viewing the last 100

tweets of a given user. The likelihood that two users had both messaged/retweeted each-other twice, or even once, in their last 100 tweets was simply too low, resulting in almost 0 edges in the graph.
As a result, I determined that a more lenient definition of "friends" was necessary, and settled on the following more general definition:

Given a user A and a user B, users A and B are friends if and only if:

user A and user B are mutual follows, AND
user A has mentioned or re-tweeted user B at least once

Using this model, I began data collection, a lengthy process that has gone on for several weeks.

I started out with a random sampling of 100-200 tweets using the Twitter Stream "Sampling" API, eliminating any user which was clearly not posting in English. This was done as a matter of practicality, as it was difficult enough to interpret the meanings and memes used by US Twitter users. It would simply have been unfeasible to also do this for non-English speaking communities with their own distinct online slang and terminology within the limited timeframe of this study.

I selected the users who posted these randomly sampled tweets to be the starting nodes in my graph. Then for each user, I recursively searched through their peers in the Twitter graph, using the "friends" and "followers" API calls from the Twitter REST API, which returns the users being followed by the current user, and the users following the current user, respectively[4]. Once I've established which of the current node's peers are "mutual follows", I inspect their last 100 tweets using the Twitter REST API's "user timeline" API call, to determine which of them meet my previously stated friend criteria [4]. Those who pass get added as nodes to the graph, with an edge created from the current user node to the newly created nodes.

This process continues recursively for 3 levels, such that in total my graph contains the seed nodes, their friends, their friends of friends, and their friends of friends of friends. This cutoff is required due to time restrictions. With a total of 4 levels (including the seed users), the number of nodes in the graph reaches the thousands very quickly. While this is not an issue in and of itself, the biggest restriction to data collection is the Twitter API itself, which allows a maximum of 350 requests per hour [4]. I use heavy caching to try to mitigate this as much as possible, but visiting and collecting data for a given user can still cause anywhere from 0 to 7 api calls to be made. The result is that the number of users that could be reached within the timeline of this study was well under 10,000, which is large enough to perform interesting and useful data analysis, but still insignificant compared with the overall Twitter landscape.

Community Identification:

Once the final data graph has been built up, I use a version of the Louvain method to very efficiently partition the graph into a set of communities of nodes. The approach is described by Blondel et al. [7] in a 2008 paper, and is based around iteratively optimizing the ratio of density of links within the communities to the density between the communities (this is also known as the modularity of the partition). There is already an existing plugin for NetworkX that implements this algorithm [8], and it has been used in this study.

Community Labeling:

After arriving at a set of communities, I iterate through each community and generate labels for it using a weighted application of Term Frequency – Inverse Document Frequency (TF-IDF). This is a frequently used text-processing metric which gives a measurement of how important a word is for a particular document in a group.

I start out by, for every user in every community, generating a dictionary of the term frequencies, one for the user's description page from Twitter, and another for the text of their last 100 tweets. This is the base data from which the labels will be generated.

Before any further processing, terms which are non-descriptive or gibberish need to be filtered out. This is particularly challenging given the inconsistent grammar and abundance of obscure slang terminology found on Twitter. I start out by normalizing all text into ASCII, to simplify the filtering process. After this, the resulting ASCII is converted to lower-case to minimize spelling differences, most types of punctuation are filtered out, and "@username" mentions are eliminated. URLs are also filtered out to try to prevent marketing and spam from having a negative impact on the study. Finally, on this cleaned up data, a stop-list is applied to filter out some common non-descriptive terms. I used the Stanford WordSift stop-list[9] as a base list, and hand-added around 20 words to improve its effectiveness.

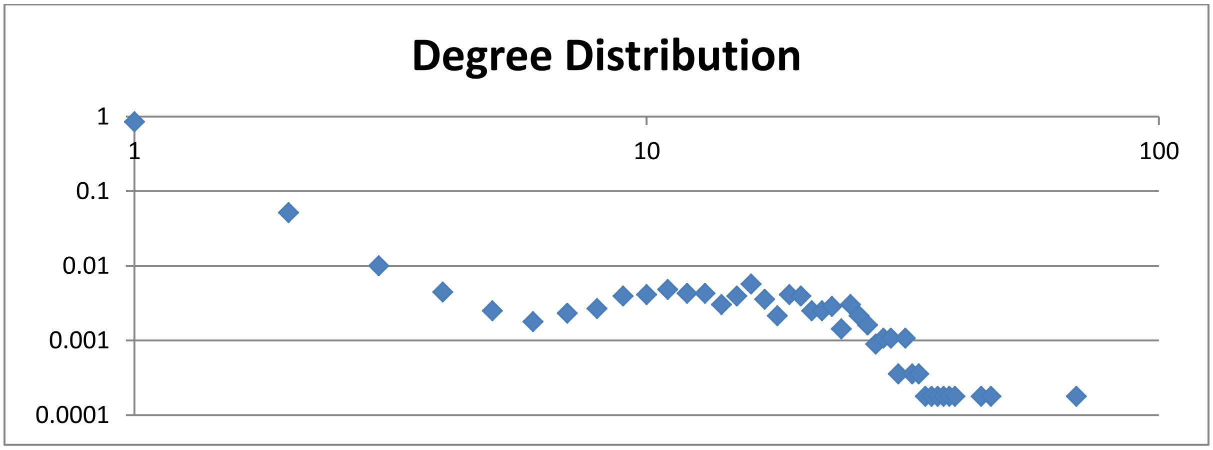
From here, a two-tiered level of TF-IDF is applied, first at the User level, and then at the Community level. Thus, I initially generate TF-IDF scores for the terms on each user, to weed out unimportant terms at a local level. Then, those scores are used as the input "term frequencies" for a second computation of TF-IDF, this time at the community level, such that each community is a document in the set of all communities. Finally, the TF-IDF scores of the data from the tweets and the data from the user descriptions are added together, with the user descriptions weighted at 75% and the tweet data weighted at 25%. This makes sense since the text from the users' descriptions is much more likely to include key data about their interests and demographics than the more general and random twitter posts. From this weighted sum, the final term scores are generated, and I have used the top 10 highest scoring terms in each community as that community's labels.

Geographic Survey:

I've also performed a basic Geographic survey to see if location plays a significant role in bringing users in Twitter communities together. This is done using user's self-reported location in their Twitter profiles, which is obtained through the Twitter API. On their own, these measurements are imprecise because many users describe their locations with variations in spelling, as well as various nicknames for their city name, while other users don't provide this information at all. However, I do a certain amount of normalization to try to mitigate this problem. Similar to Community Labeling, I start out by converting all text to lower-case ASCII and parsing out most punctuation. Then, I use the Cities-1000 dataset from GeoNames [10], which provides a listing of all cities globally with a population over 1000, as well as common nick-names for each city. I convert this data into a dictionary and use it as a normalization table, converting any city nicknames from my data into the full city-name from the Cities-1000 dataset. Finally, with the normalized data, I'm able to generate a frequency map for each community, showing how many users in each community are located in each location.

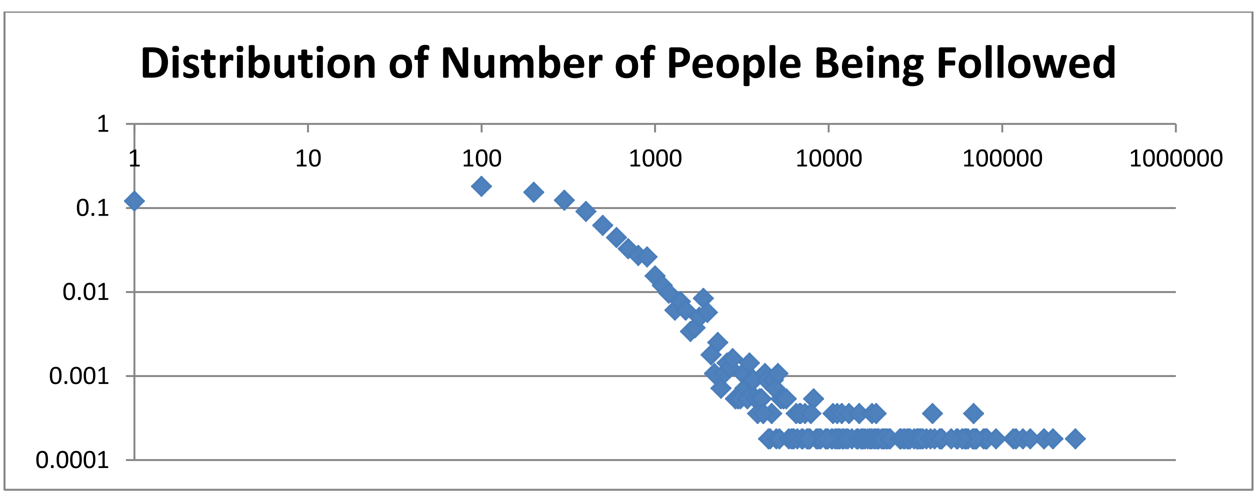
3. Results

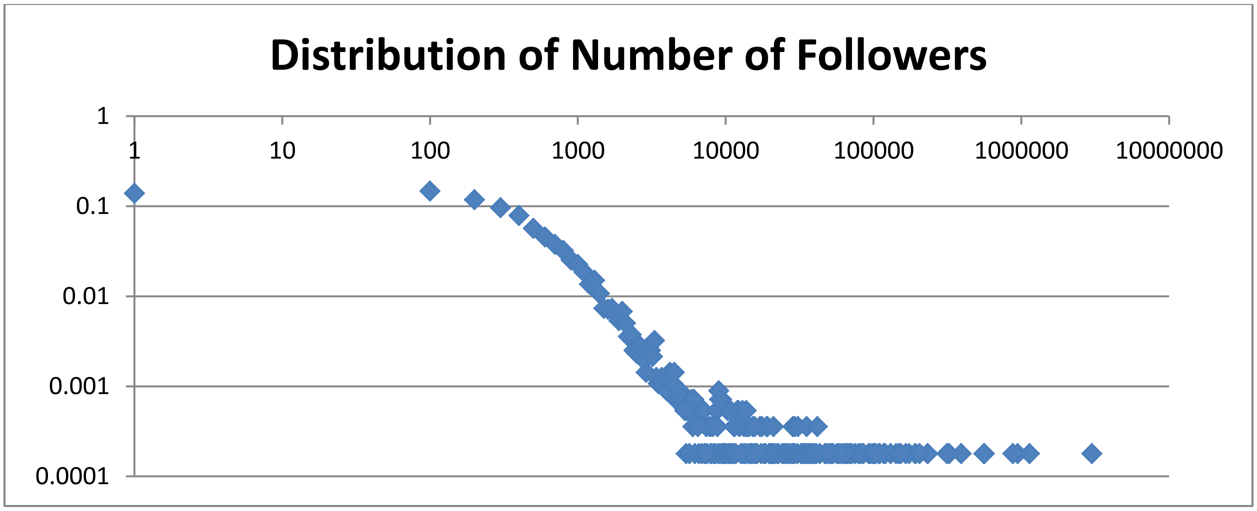
Overall, my final data graph consisted of 5593 user nodes, with an average of 2.36 friend degrees. Each user had a median of 335 people they were following, and 399 followers (averages are not very useful here due to a few massive celebrity account outliers within a relatively small dataset). The degree distribution for the overall Twitter graph can be seen below:



As you can see, no user had over 100 "friends", under my restricted definition, with the large majority having only one. This is an artifact of the fact that we only do a partial traversal of the Twitter graph, with nearly every node at the "horizon" at which we stop, having only one friend.

For reference, the Follower and Following distributions, which mirror each-other very closely, are also provided below. As can be seen in both cases, the majority of users have between 0 and 1000 followers, and are following between 0 and 1000 other users, with a quick exponential drop-off afterward.





The community generation effort distributed the 5593 users into a best partition of 47 distinct communities. Below, I provide a detailed table of the top key-term labels for these 47 communities, along with the score for each term (higher means a greater importance for that community). I've also listed the number of users in each community.

As a whole, the combined approach of community identification and automatic labeling appears to be a surprisingly effective way of determining what kinds of unique topics have meaning for users in these communities. For instance, we can see that Community 1 is clearly very interested in marijuana, with labels such as "#smoketothat", "#marijuana", "#greenlightsociety", and "#hempembassyforum". Alternatively, community 16 appears to have Christian leanings, with top labels such as "#saturday_sabbath", "Sabbath", and "Christians". There are also several video-game related communities, such as numbers 2, 3, and 39. This approach also allows us to see interesting correlations between discussion topics. For instance, community 2 has clear videogame leanings, with labels referencing the latest Elder Scrolls game, Skyrim, as well as the latest Zelda game, Skyward Sword. At the same time, we see this community has an interest in Atheism, with both "#athiest" and "#atheism" as labels, and it also has "anime" as the 3rd most important term. If we were to suppose that characteristics of Twitter users reflect their real life selves, such data would suggest that there is at least some overlap between the gamer community, the anime community, and the atheist community, at least for this small group of people. This is by no means the only example of this. If we look at Community 23, these users would seem to have an interest in the track sport, as evidenced by key labels "#trackgoals" and "#tracklife", while also having an interest in or at least knowledge about the lakers basketball team, as evidenced by the "#nowplaysforthelakers" key label. The fact that the high scoring key-terms can be this readable and descriptive of these communities, shows that this approach of community identification and community labeling is indeed a valid, and very powerful tool in finding the interests and characteristic of online users.

Community 0 (261 nodes)		Community 1 (165 nodes)		Community 2 (263 nodes)	
#bijou	1199.76	#smoketothat	2767.69	#netheads	943.96
georgetown	955.75	#mmot	1631.67	Skyward	891.30
observing	736.90	#marijuana	1144.81	Anime	880.84
#dontevengetmestarted	728.42	#mbmbitch	845.32	Bliptv	772.33
winds	700.85	#uallreadyknow	741.50	Sonic	700.59
#yb	621.30	#greenlightsociety	471.81	#atheist	689.94
pearland	544.58	#taf	430.91	Skyrim	618.45
dawson	529.81	hempembassyforum	393.17	#atheism	615.70
reporting	484.95	#shrim	339.67	Zelda	577.76
mph	472.00	sf2	322.39	Sword	552.40
Community 3 (289 nodes)		Community 4 (209 nodes)		Community 5 (252 nodes)	
skyrim	1996.63	sojesuscristosalva	1151.85	#planogirlprobz	1681.84
realm	1244.29	jesuscristoteama	822.75	Felo	1639.26
gaming	1152.91	d81	493.65	#smtx	1554.11
#skyrim	973.38	#detroitdreamz	473.08	Marcos	1338.10
zelda	719.07	buuk	431.94	#machida	1043.17
xenoblade	571.36	pandemonium	381.72	Yoko	1042.41
#usab2011	523.60	#badsanta	269.85	#1000waystodie	979.30
360	518.34	#lostluxuries	267.39	#beforethefame	872.82
anime	509.77	#waystomakeanewburggirlhappy	267.39	#jones	855.36
#workshopsf	501.78	12-23	264.59	#txst	768.08
Community 6 (160 nodes)		Community 7 (208 nodes)		Community 8 (157 nodes)	
kid-ro	1328.73	lb-	1048.07	#yikess	830.05
#wavygotit	1055.17	aztecnical	965.86	Nique	811.80
#joebudden	849.18	produceddirected	945.31	Ltrbb	661.89
banger!!!!	849.18	#justfortherecord	739.81	Dietgt	642.42
follow==gt	468.96	krs-one	734.32	#teamufb	593.24
#25	329.79	#freemumia	602.14	mista	541.20
#imnotevenkidding	273.56	#graffiti	543.40	#foetaughtme	467.22
isu	215.68	#zimhiphopawards	534.31	#codewordsforsex	414.59
niko	198.49	#pdxmusic	513.76	#dec6	408.81
shoutoutgt	175.86	#streetart	505.52	#codewordforsex	344.41
Community 9 (86 nodes)		Community 10 (119 nodes)		Community 11 (227 nodes)	
#okeboyz	583.10	417pm#norunnerup	809.61	upgettin	1482.97
mazi	351.56	11-30-11	699.21	sns	1198.86
adriannas	296.87	psa!!	663.86	#boyparty	814.59
2many	220.61	$h!t	607.21	#youngflyshit	731.04
163rd	210.94	astonish	591.75	#tcas	692.88
754	210.94	#theblooddiamondtape	543.16	#queenroe	543.06
#rtb	205.80	#fantasy	518.10	#thecrownaintsafe	537.18
reese-	205.80	$#!t	473.40	#bruins	355.08

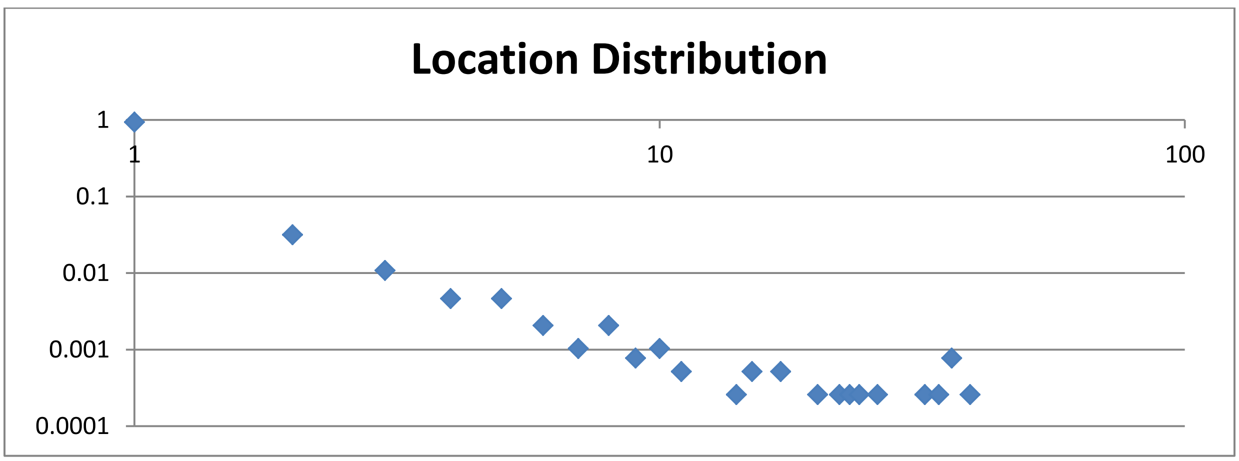
durklil	205.80	publiquei	460.01	#sammyadams	342.53
#tiandtiny	196.87	12054	441.61	holyoke	313.30
Community 12 (212 nodes)		Community 13 (253 nodes)		Community 14 (106 nodes)	
herestothekids	1918.00	lhh	1156.16	#swerge	1095.25
#keepsitreal	928.06	mizzou	795.73	#teritweet	551.52
#atb	903.57	#stopdayback	646.35	#swergin	412.96
#m-u-s-t-f-o-l-l-o-w	886.82	obs	639.13	99cents	359.10
#raam	625.70	unbrelievable_	404.78	#repyocitystate	305.23
#vip	541.14	rollins	395.86	hideaway	251.37
#luv	540.11	#certifiedbruhz	362.17	lima	215.83
#weekend	524.23	roo	299.94	#faf	205.31
#special	507.32	#hue	298.26	indie	181.62
follow-	474.34	#srbym	276.96	#thekings	179.55
Community 15 (77 nodes)		Community 16 (120 nodes)		Community 17 (236 nodes)	
#rackcity	786.04	#saturday_sabbath	1456.17	jino	1514.63
#whatsonyourhanger	685.70	sabbath	474.73	#openmicjam	1262.19
#lookdujour	518.45	#itsimpossibletobeathugand	461.05	4731	1262.19
#blamedavidstern	351.21	braxtons!	442.38	#the1percenter	946.65
wallace	341.95	2025609934	368.65	#gtp	820.43
forget!!	286.85	rcn	368.65	#buyart	736.28
twipic	260.55	christians	305.18	#becultured	736.28
#reasonstoloveawoman	250.99	tns	294.92	#4731gallery	715.24
#transparentworship	250.86	#ibizasaturdays	294.92	#ilovefridays	694.21
tucker	241.81	#vs	287.17	#occupythebottega	694.21
Community 18 (201 nodes)		Community 19 (206 nodes)		Community 20 (44 nodes)	
owey	1776.41	#sm4cu	2707.73	#mantipoftheday	233.12
teamwe	1674.32	#wineparty	1456.43	wassam	116.56
donksdimesdiamonds	1653.90	#headstart	1128.22	nahs	101.99
dyerk	1633.48	#girlsmediachat	964.12	hometeam	101.99
basketballz	1633.48	#mdmq	807.37	dissy	87.42
#leak	1305.92	#santacon	605.52	#tellthegametostfuforareactionday	72.85
#stackorstarve	1289.18	#ifyoufromthesouth	533.34	#wordofwisdumb	72.85
chasendough	948.49	#foodiechat	533.34	lt----#nf	72.85
#stackorstarveeverythingz	837.16	#rhobh	485.70	bsf	72.85
consulting	828.10	fayetteville	483.77	hotttt	59.73
Community 21 (53 nodes)		Community 22 (65 nodes)		Community 23 (165 nodes)	
milledgeville	580.88	#swayze	851.82	#trackgoals	589.76
sparta	303.25	#kreganddezshow	594.66	#ltr	499.70
samos	214.01	#wheniwaslittle	187.07	centro	371.35
#30dayselflovechallenge	214.01	tweetiers	160.72	#nowplaysforthelakers	353.86

#mlk	200.55	#hoodmemories	154.67	#tracklife	314.54
#classiccomedy	198.72	youulovebri	128.58	#ccu	294.88
#mantra	168.15	drillllll	128.58	oceangang	294.88
#morningdevotion	168.15	skitzz	112.50	#maw!	294.88
graystone	152.86	#ren	112.50	ashland#sicklife	274.03
#select	152.86	#iconfess	112.50	playmatekey	255.56
Community 24 (7 nodes)		Community 25 (183 nodes)		Community 26 (133 nodes)	
korra	80.11	rayla	501.43	#webenfresh	546.03
comic-con	43.00	gmo	487.36	#ailitemusicgang	527.20
showcase	39.22	#lovegivenforlovereciev ed	378.27	#cirquedeexquisite2012	336.40
panel	37.74	#teame#mentionme	340.97	weeze	308.78
atla	36.86	#cbnrentrunway	340.97	#wingsup	301.26
kung	31.88	#intuned	340.97	#nye2012	295.86
#korra	30.72	#codewordsforsex	318.61	#amg	293.34
sdcc	22.48	#nonightsoff	260.74	chicago!!	280.02
#felizdiadelmusico	22.48	foc-	246.70	skooda	277.90
romney	21.08	prequels	243.68	#typesex	269.12
Community 27 (16 nodes)		Community 28 (86 nodes)		Community 29 (41 nodes)	
ceefax	74.72	#viewhiphop	898.28	stlaz	1143.82
indexing	64.05	lt---view	463.05	#smoke	640.48
#blackmirror	53.40	lt----view	407.81	#news	632.40
#pubnow	53.37	spot!!!	309.37	#musicvideo	574.15
waitrose	43.77	streetz	279.47	#otg	471.83
brighton	43.77	problemchild-prime	274.40	#checkitout	443.23
#lastfm	36.96	upscale	269.64	#youtube	420.80
aphex	35.01	ctfu	255.81	#ssscanada	400.34
#howitsmade	32.02	murda	236.62	#bitches	343.98
pheasant	32.02	superi	222.95	#hoes	334.09
Community 30 (27 nodes)		Community 31 (114 nodes)		Community 32 (171 nodes)	
nox	456.82	#atheist	560.37	#10day	3881.66
#venue	190.34	#israelhates	513.45	#happybirthdayjayz!	1635.42
danes	154.29	arbol	510.58	chara	1466.56
#realpeople	126.89	#atheism	448.57	#kingshit	847.27
grille	110.78	alto	418.66	#savemoney	741.34
#roxy	104.05	floaty	401.17	savemoney	513.66
#thehorn	101.52	#littleknownndaaprovisi ons	364.70	#windows	472.89
lumen	99.21	#madonnasuperbowl	310.00	#debbie	453.19
nox!	88.83	irl	297.49	debbie	394.08
#timeoutdallas	88.83	orrery	273.53	mee-ae	378.47
Community 33 (102 nodes)		Community 34 (79 nodes)		Community 35 (87 nodes)	
detat	783.50	narly	1127.14	#teamchaosusa	2100.73
coup	585.39	#superb	958.91	#ifollowall	890.61

#ss3	391.75	#clevelandmusic	908.44	#500aday	858.35
idiota	356.14	#download-n-listen	857.97	pressn	835.59
#nowfollowinq	338.33	#cntr	432.82	#teamfollowback	775.64
#rolextalk	248.22	tezo	336.46	#justinbieber	775.44
#jdr	204.41	#slanderteam	206.92	#mustwatch!	739.36
#lifestyletuesdays	175.21	==gt	201.83	#ifollowback	649.66
#capricorn	167.03	rt@ryanpdotcom	185.05	#tfb	557.67
#fitb	158.47	#thanks	155.38	#instantfollow	552.77
Community 36 (45 nodes)		Community 37 (73 nodes)		Community 38 (15 nodes)	
#livingood	481.81	#raw	182.33	dayza	83.41
#coldcorner2	249.16	depauw	148.67	brionna	83.41
248540-0150	175.87	seq	132.15	#thingsnottosayonafirstdate	41.71
j-culli	161.22	numadosmil	132.15	shaela	41.71
#hoodratanthems	102.59	#codewordsforsex	129.76	#youarenotmachines	31.28
dylan	58.91	dalvin	121.90	gdmorning	31.28
#bezzlegang	58.62	ainte	115.63	gdnight	31.28
#thingsthatblackpeopledo	51.15	theo	111.29	#1rstalbumbetter	31.28
#12dayclassic	48.07	#flyanddope	108.36	triv	20.85
#keywanefor2012xxlfreshmanproducer	48.07	chiquira	94.81	#usr	20.85
Community 39 (32 nodes)		Community 40 (69 nodes)		Community 41 (36 nodes)	
#leagueoflegends	196.40	zelda	393.15	hiram	256.36
dota2	162.11	warwick	349.51	rglnd	179.36
rochester	128.09	skyward	330.57	#vim	165.56
gh0stick	120.09	pso	277.13	#pray4hiram	141.27
#occupyla	116.60	homestuck	218.21	#teamfuckit	137.97
#lostboysthethirst	106.75	#creeptweets	211.93	raunchy	133.25
lithium	106.75	comics	207.94	mafxckas	96.58
totton	93.41	comic	202.05	#pray4hiram!	82.78
minecraft	80.20	kart	195.90	#latinamob	82.78
#nerdisms	80.06	mlp	189.75	#pray4himram!	68.99
Community 42 (74 nodes)		Community 43 (12 nodes)		Community 44 (25 nodes)	
#ludingtonmemories	381.14	kamen	78.45	#movieacademia	136.32
#chrome	215.43	breton	42.86	maar	132.11
rfk~streets	198.86	neverland	42.59	#ghcomm	123.93
us~	198.86	dunmer	34.19	#shakeshake	123.93
hoodstarz~	198.86	#skyrimproblems	28.70	det	117.05
#bethere	130.27	da2	27.35	een	107.52
rfk	116.00	#balls	27.35	voor	106.28
007	106.55	rider	26.78	ubisoft	88.57
1075	106.31	korra	25.57	skyrim	86.71
#ludington	99.43	comic	25.05	3ds	84.14

Community 45 (20 nodes)		Community 46 (12 nodes)	
greeley	66.20	#toyshow	101.98
#walewednesday	56.75	soz	62.76
#famu	49.33	limerick	54.70
famu	44.99	bbz	47.86
foco	37.83	awks	41.02
#imasexyassguy	34.60	saigon	38.27
#dropthatbitch	34.60	fwends	38.27
overload!!!!!!	34.60	panto	38.27
#probablydoesnttakem uch	34.60	glasgow	34.19
#disappointed	34.23	centrespace	28.70

In addition to community labeling, I also generated a frequency list of geographic locations for each community. However, what quickly became surprisingly apparent was how little commonality there was in terms of location between the users in each community, as well as globally. Below is a global location distribution for all 5593 users.



As you can see, the vast majority of locations only had 1 user self-registering them-selves there. Part of this is a data-analysis issue, because of the huge variety in location spellings and descriptions. Though I perform a significant amount of normalization, there are still several cases where two users who are in essence living in the same city will be treated as users in separate cities. That being said, even factoring this in, location plays a much smaller role in bringing Twitter users together than I initially expected. In fact, it does not appear to be a very significant factor at all.

4. Conclusion

From the results of this study, it seems clear to me that the combined approach of automated community detection and automated community labeling is not only possible, but a very powerful tool in learning about the interests and habits of Twitter users. While this particular study has focused on Twitter users, I see no reason this same approach cannot be applied to other online Social groups such as Facebook and MySpace. At the same time, this study also indicates that location is not a very significant factor in tying Twitter users together, though future studies applying more in-depth location normalization techniques can make a more definite statement on the level of significance location plays. It should also be remembered that this study has been based on a very limited sample size of the actual Twitter graph, and that only the last 100 tweets of each user in the study have been checked. A longer running and wider-spanning future study would undoubtedly be able to find more conclusive correlations between the interests of users within the Twitter social graph. That being said, this study has shown that this combined method is effective, and I hope to test it on larger and more diverse datasets in the future.

5. References

1. H. Tsukayama. Twitter hits 100 million active users, *September 9, 2011*. *http://www.washingtonpost.com/blogs/faster-forward/post/twitter-hits-100-million-active-users/2011/09/09/gIQAs9QjEK_blog.html*

2. J. M. Pujol, V. Erramilli, and P. Rodriguez, Divide and Conquer: Partitioning Online Social Networks, *May 29, 2009. Available at ARXIV: http://arxiv.org/PS_cache/arxiv/pdf/0905/0905.4918v1.pdf*

3. T. Kalafatis, Twitter Analytics: Cluster Analysis reveals similar Twitter Users, *May 2009. http://lifeanalytics.blogspot.com/2009/05/twitter-analytics-cluster-analysis.html*

4. Twitter API documentation. *https://dev.twitter.com/docs*

5. NetworkX graph library for Python. *http://networkx.lanl.gov/*

6. B. A. Huberman, D. M. Romero and F. Wu, Social Networks that Matter: Twitter Under the Microscope, *December 5, 2008. Available at SSRN: http://ssrn.com/abstract=1313405*

7. V. D Blondel, J.L. Guillaume, R. Lambiotte, R. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment, *2008, P10008 (12pp)*. Available at ARXIV: http://arxiv.org/PS_cache/arxiv/pdf/0803/0803.0476v2.pdf

8. Thomas Aynaud, Community detection for NetworkX. http://perso.crans.org/aynaud/communities/

9. Stanford WordSift list of StopWords. *http://www.wordsift.com/wordlists*

10. GeoNames Geographical Database. http://download.geonames.org/export/dump/