Rinal Chheda

Analyzing Twitter Networks

1.204 Course Project

Introduction

Twitter is one of the newer social networking sites started in 2006. It is different from most social networking sites in that it is used to spread news and pass information more than it is used for interacting with one another. Publicity and spread of information are a major reason for using twitter. As I am an active twitter user, analyzing this data was very interesting for me to see how connections on the twitter network function.

Data

Twitter data is available at several places on the internet. Some of the sources I considered are as follows:

1) Stanford twitter data
    a. 476 million tweets with over 17 million edges
    b. Higgs twitter data – over 456,000 nodes and 14 million edges, retweet, reply and mention network
2) Gephi twitter dataset – only has twitter mentions and retweets of some part of the Twitter network
3) Twitter API
4) Personal data

A major hindrance in using the first dataset was its size. The size of the network in the Stanford dataset is not supported by gephi on a laptop. The gephi dataset did not have the data of followers and followees. In case of personal data, it was difficult to get data from more than five friends. Besides it was a small network and would have not given accurate results. It was important to get data from random users, hence I used the twitter API to extract data of 20 users. The data used was –

- For 20 random active users from Boston area (no celebrities)
    - Their follower count – number of followers the user has
    - Their followee count – number of people the user follows
    - Their tweet count – number of tweets by the user
    - Direct and indirect tweets by the user from two months
- Over 15,000 random users from across the globe (may include inactive accounts and celebrities)
    - Follower count
    - Followee count
    - Tweet count

Previous work

There has been a lot of work on analyzing twitter networks. Huberman et al[1] have studied social interactions on twitter to find that there is a hidden network of connections underlying the declared set of followees and followers. They plotted the follower count, followee count and tweet count of over 300,000 users to find patterns. They also tried to find the underlying actual friend count for users in the twitter network. A user was considered an actual friend if there were more than 2 interactions directed towards that user. They found that the number of posts (tweets) increase initially as number

[1] Huberman, Romero and Wu; Social networks that matter: Twitter under the microscope

of followers increases but it eventually saturates (figure 1). Their next finding was that the number of posts (tweets) increases as the number of friends increases without saturating (figure 2). The other finding was that the number of friends saturates while the number of followees keeps increasing due to the minimal effort required to add a new followee (figure 3).

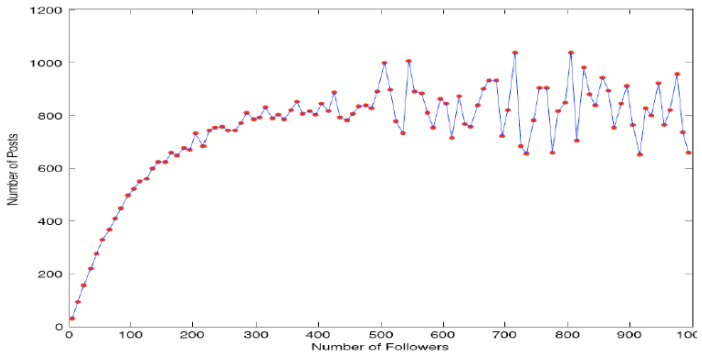


Figure 1 Huberman's findings: Number of posts vs number of followers

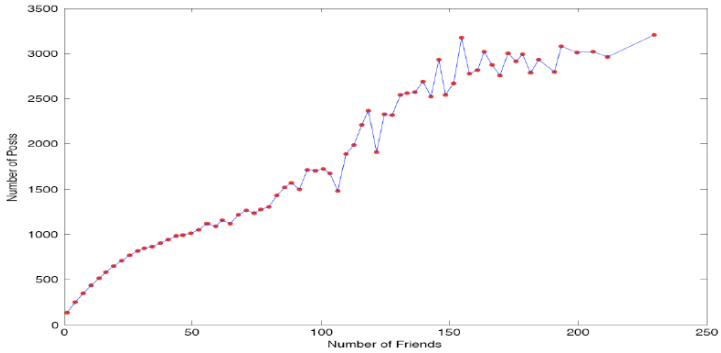


Figure 2 Huberman's findings: Number of posts vs number of friends

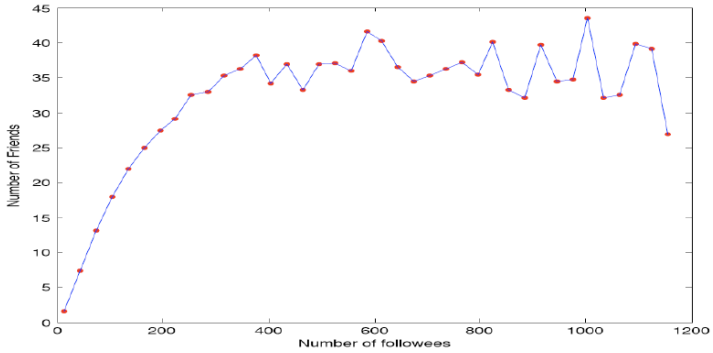


Figure 3 Huberman's findings: Number of friends vs number of followees

In conclusion, Huberman et al find that even when using a very weak definition of friends (2 directed posts), Twitter users have a very small number of friends compared to the number of followers or followees they declare. The paper does not mention Dunbar number or quantify the an average number of actual friends within a person's twitter network.

In the next sections…

I have presented three findings from the twitter data I had. In the first part I have verified some of Huberman's findings. I have shown patterns between the follower count and followee count and tried to relate them with each other as well as with the tweet count of the users. In the second part, I will

discuss about how to effectively spread news or information on twitter using organizational accounts. The third part is to study the actual friends of a person compared to the perceived friends (a network within a network) by studying the number of direct tweets of a user with certain people. I have proved that Dunbar's theory holds for twitter networks as well.

About Twitter

Twitter is a directed network. It allows users to post short messages (less than 140 characters) that can be read by other twitter users. Users declare people they are interested in following (followees). Thus for each users, there is a set of followers (people who follow the user) and a set of followees (people whom the user follows). Users can post direct or indirect tweets. Tweets are direct when the user addresses them to a specific person by using the character '@' before their twitter name. Indirect tweets are for anyone who cares to read them.In this report, the number of followers a person has will be called follower count. The number of people a person follows will be called as followee count.

Follower Followee

Methods and Results:

Part 1: Analyzing patterns in user networks

In-degree: follower count

I plotted the in-degree histogram (number of followers of any user) of over 15,000 users. As expected, a high number of users had very few followers. These users are either new to twitter or have little activity on their accounts. Very few accounts have a huge number of followers. These accounts typically belong to famous twitter users or celebrities. The graphs (figure 4) quickly tapers off for number of followers greater than 2000. The average in-degree of all the 15,000 users is 6877 since two users have as many as 300,000 followers. If we remove these outliers from the data, then the average in-degree is 587.

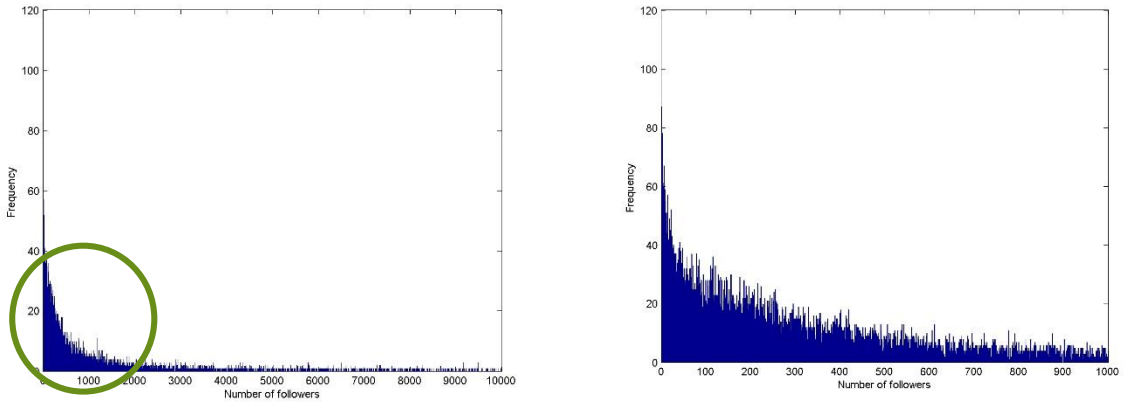


Figure 4 Histogram of follower count of 15,000 user data showing users with 10,000 followers or less. The circled part is zoomed in in the right-side graph.

Out-degree: followee count

Next I plotted the followee count i.e. the number of people a user follows. Again, the results were similar to what I expected. Most people follow less than 3000 users. The number of users that follow

about 2000 people is high. A few users follow over 10,000 people. I would assume these are spam accounts that follow most users they come across. There is a limit to how much information a user can actually get from his followers and hence it does not make sense to have too many followees. So I did not expect many users to have over 3000 followees. The following graphs (figure 5) show the distribution of number of people that users follow (followees). The average out-degree for all users is 764. Considering only users with less than 1000 followees, the average out-degree falls to 575.

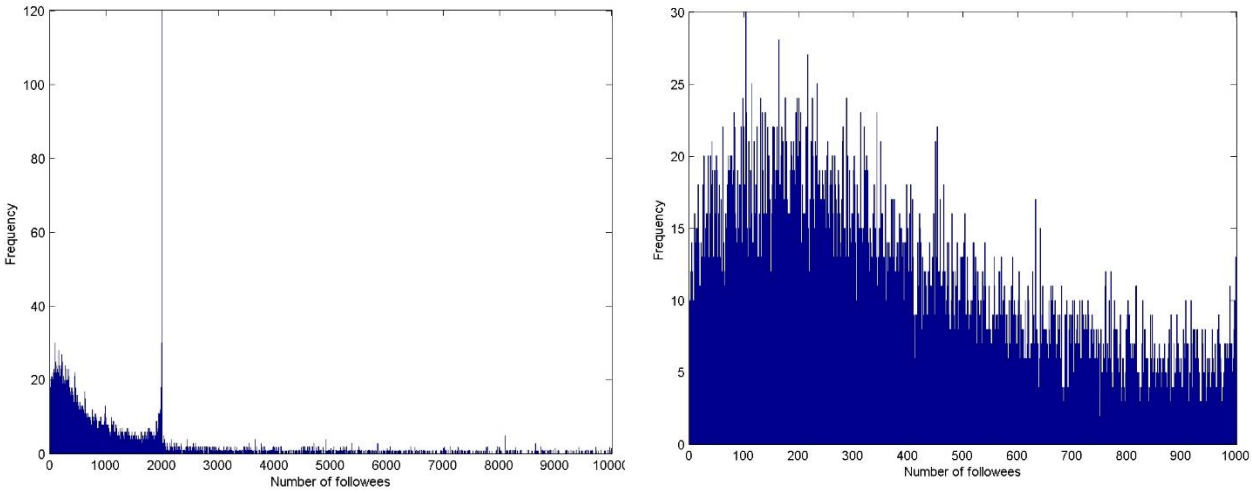


Figure 5 Distribution of the number of followees. The right side graph shows a zoomed in version of the same graph.

Followee count vs follower count

I plotted the followee count vs follower count for all the 15,000 users on a semilogy scale and binned the data to see a clear trend. The graph for followee count vs follower count is as I expected. Number of followers of a user can increase indefinitely but there is a limit to the information a user can get from his followees and hence I expected that the number of followees would eventually become constant even though the number of followers increases. The following graphs (figure 6) show the observed trend.

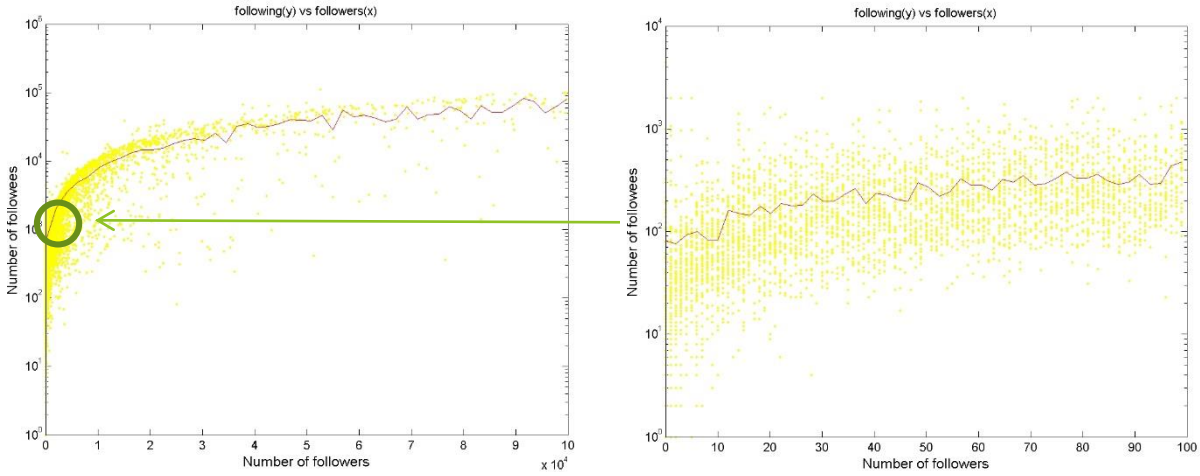


Figure 6 Followee count vs follower count

Number of tweets vs follower count

From the graphs below (figure 7), it can be seen that more a person tweets, the more followers he or she has. However there is a limit to the number of tweets a person can make in a day or a week.

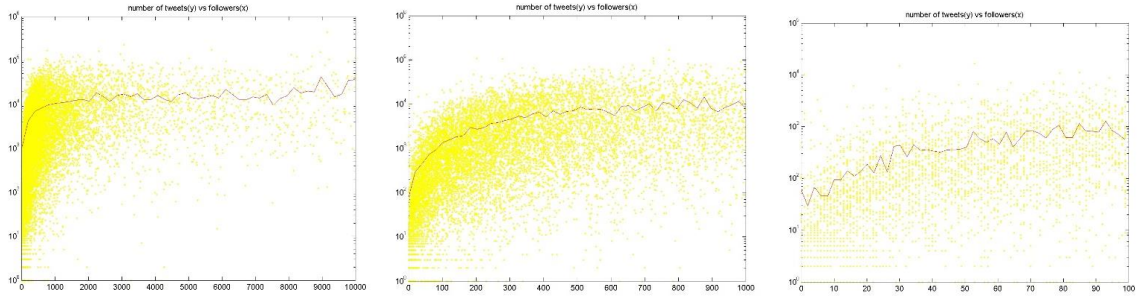


Figure 7 Number of tweets vs follower count

I also plotted the data on a linear scale (figure 8). It showed some clear distinction of different types of twitter users – regular users, new or inactive users, popular twitter users and celebrities based on where their point was located in the graph.

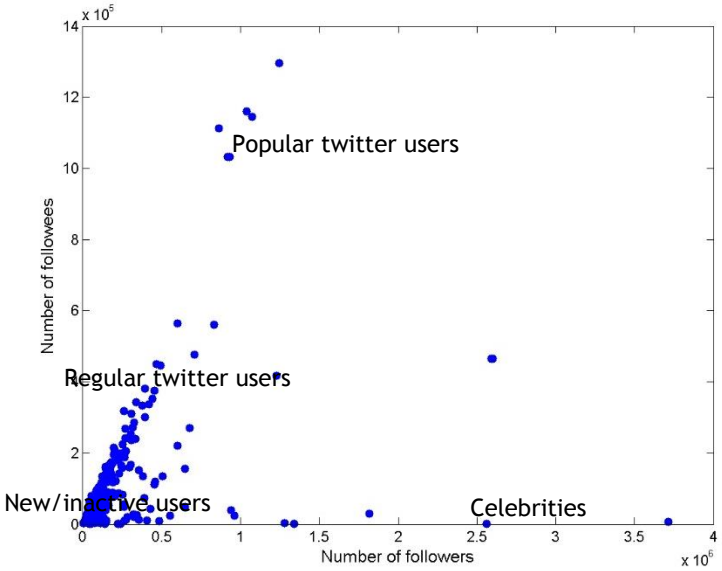


Figure 8 Identifying types of users from the followee and follower count

Part 2: Twitter network

In this section, I analyze the data from 20 random users in the Boston area. The following image (figure 9) shows the users and their followee network. None of the 20 users follow each other. Also, they have been arranged in order of decreasing followee count (clockwise).

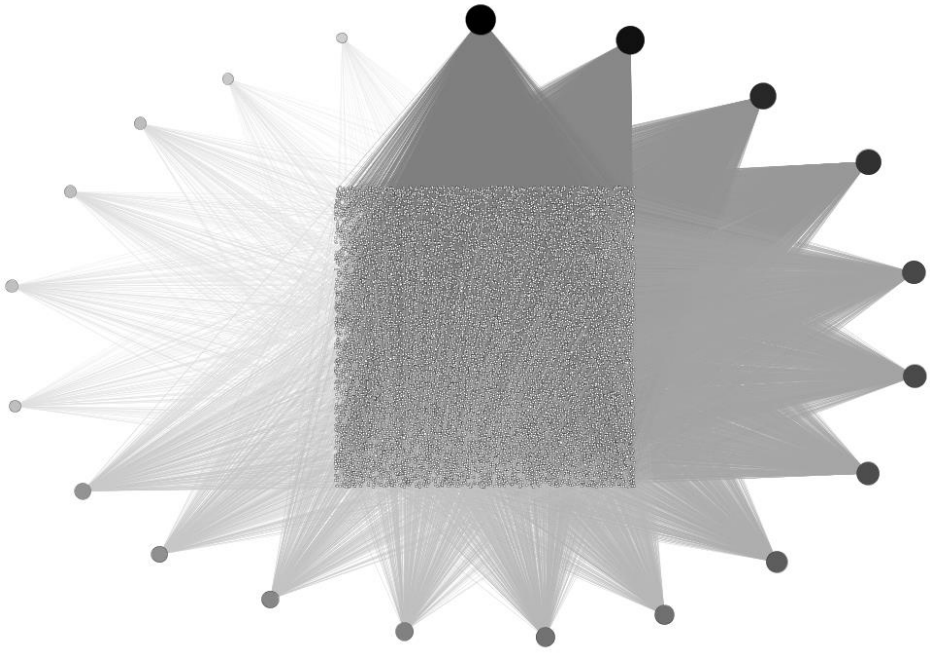


Figure 9 Followees of the 20 random users

From the data of the users they follow, I attempted to see if there were common users that most of them followed and who they were. This could have important implications about how a set of random users follow different twitter accounts and how this can be used to spread information in a network. As expected, 6 users out of the 20 followed one common account. There were other accounts which were followed by 5 or less users. This is shown in the image below (figure 10).

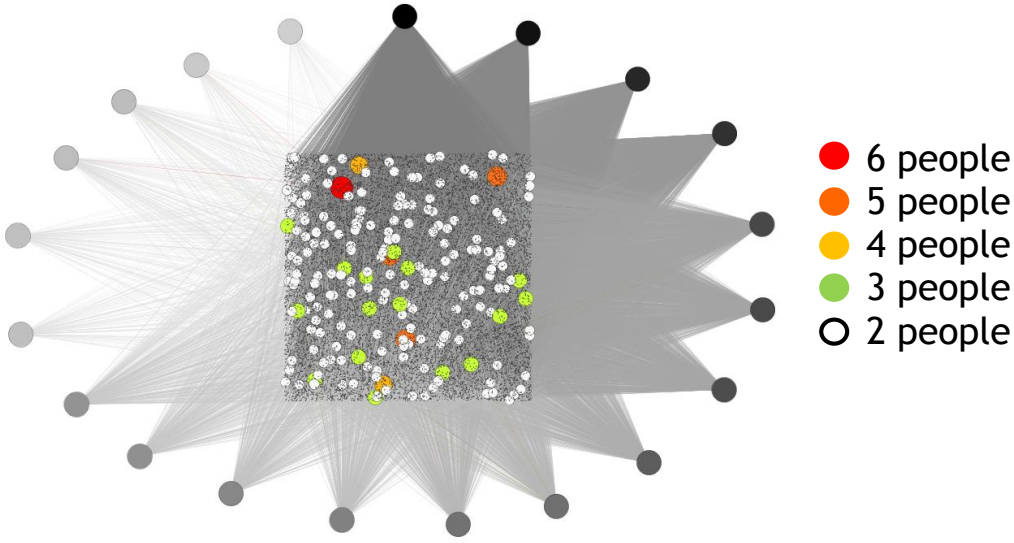


Figure 10 Common accounts followed by few of the 20 users

I separated out these accounts to understand which users these were. I expected that these accounts will be some celebrity accounts like President Barack Obama or some Hollywood actors or sports persons that most users tend to follow. The results were not what I expected, which make this finding very interesting. The accounts that were common among the users were organization accounts or event

accounts and not actual person accounts.

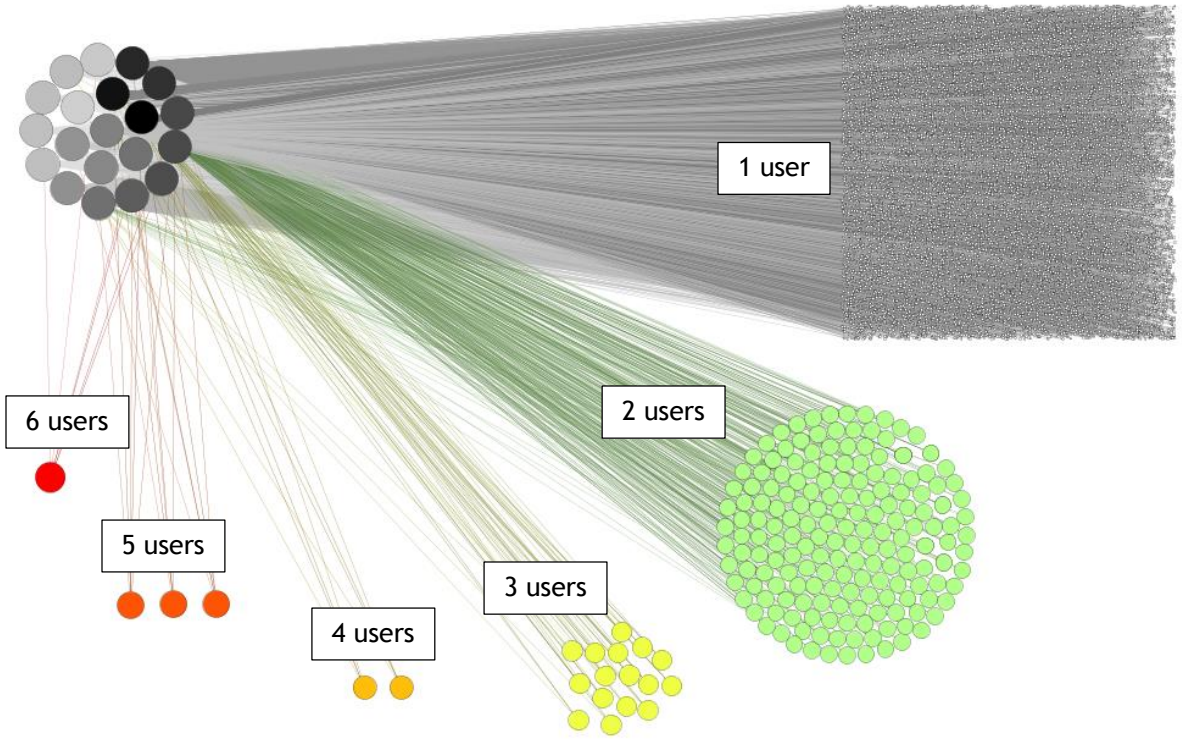


Figure 11 Identifying the common users followed

Accounts such as SocialInBoston, hiddenboston, tdgarden, HOBBoston, globeevents are social groups or organizatins that tweet about daily happenings in the Boston area and news related to the city. This has important implications for publicity, marketing, business promotions and spreading other information. While it is common for people to ask celebrities for retweets of information, if a user wants to reach the local people, then they should focus more on getting a retweet from such local organizations.

Part 3: Real friends vs perceived friends

As it is known, the number of followers or followees in a person's network are not their actual friends. Often, the number of actual friends is far fewer than the people users are connected to on social networks. This has been theorized by Dunbar. According to his theory, there is a limit to the number of people with whom one can maintain stable social relationships. This number is usually between 100 and 230 for offline social network. I found Dunbar's number for the twitter network formed from 20 users. Although this isn't representative of the whole network, it was a good exercise to see how close the number is to the range predicted by Dunbar.

I checked the theory on the data of tweets from the 20 users. I used three different ways to define real friends – users who had exchanged at least 20 direct tweets (strong friendship), at least 10 direct tweets and at least 1 direct tweet (very weak friendship).

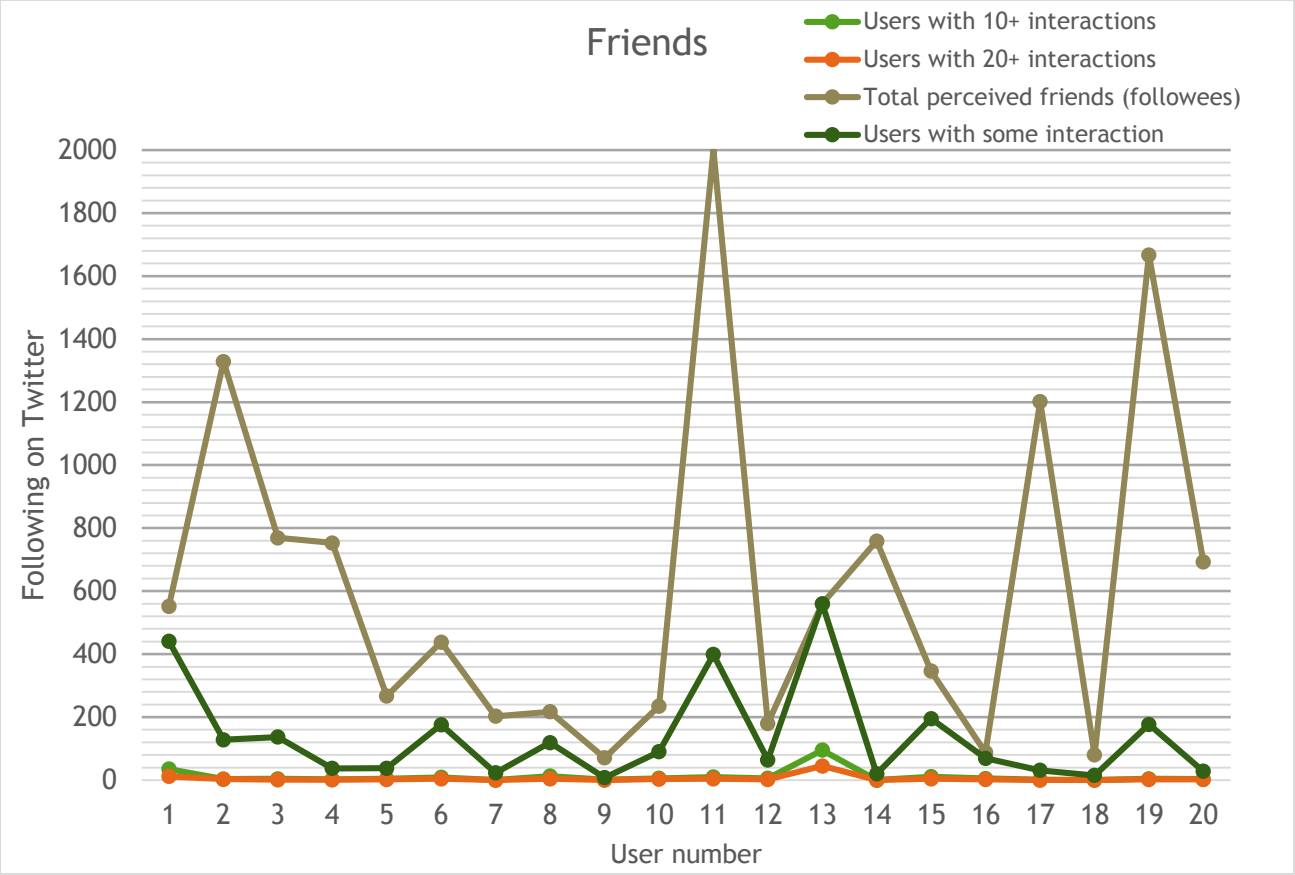


Figure 12 Actual friends vs perceived friends

The users with 10+ and 20+ exchanged interactions almost overlap each other and hence the 10+ interaction graph is not clearly visible. It is seen clearly for user number 1 and 13.

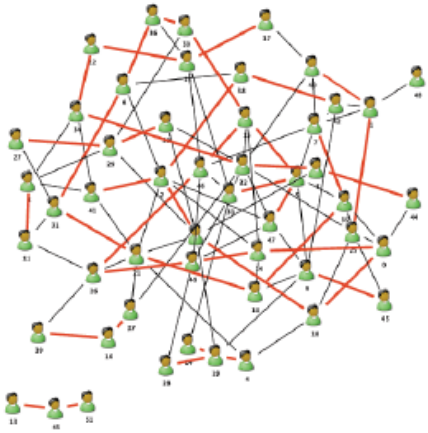
From his graph (figure 12), it is clear that actual friends are nowhere close to number of followees. The average number of followees of these 20 users is 620. The average actual friends of these users are:

- 4 friends for friends defined with 20+ interactions/direct tweet
- 11 friends for friends defined with 10+ interactions/direct tweet
- 137 friends for friends defined with 1+ interaction/direct tweet

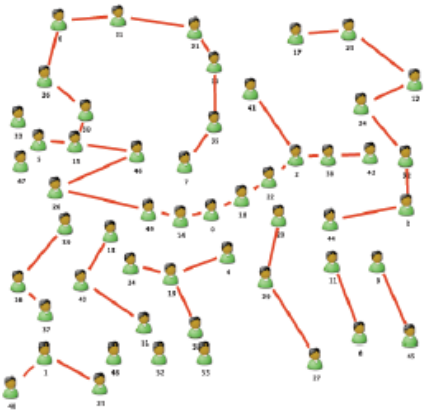
Thus twitter networks also follow Dunbar's theory, though, Dunbar's number for twitter networks is significantly low.

Conclusions and future work

The results show that social organizational accounts should be most preferred for spreading information, getting visibility or promoting businesses. It can also be concluded that like most social networks, the actual friends a user has are much fewer than the followee count of the user. This was also proven by Huberman et al in their paper. An illustration from the paper is shown below:



(a) All links are declared followees and the red links are actual friends.



(b) After removing the black links and reorganizing the network look simpler than before. This is the hidden network that matters the most.

Lack of data was a major drawback for further in-depth analysis. Having data like the retweets and direct tweets of more than 20 users would help to make a more solid case. Also, there were limitations of using the huge data on Gephi as it would hang several times while visualizing data for the 15,000 users.

A link between any two people does not necessarily imply an interaction between them. Most of the links declared within Twitter were meaningless from an interaction point of view. There is a need to find the hidden social network so that twitter networks can be used more effectively. Currently, over 40% of tweets are pointless babbles, 37% are conversational and 5.9% are self-promotional[2].
Twitter can also be used as a thermostat of the mood of the people by analyzing their tweets and the trending topics. There is a vast amount of research that can be done using twitter networks.

[2] Wikipedia, www.wikipedia.org/wiki/Twitter