The Humanity of Twitter Retweets and Mentions

Xiao Yun Chang, Joshua Fabian
Massachusetts Institute of Technology - 1.041, Network Science
18 May 2014

Introduction

Twitter is a microblogging service very popular throughout the world, with more than 200 million monthly active users expressing messages to the world in less than 140 characters per "tweet."

The shear size of the Twitter network gives us many opportunities and different avenues for study. Instead of simply examining the basic network of followers, we will consider the forwarding of messages from account to account. Based on the "Twitter Mentions and Retweets" data provided by the Gephi we find that this Twitter feature yields a small-world network with several scale-free properties.[1]

Motivation and previous work

While much research has been conducted on the following of Twitter users — generating directed networks based on who is following who — little work has gone into the investigation of the networks that could be formed when reviewing the mentioning of users within tweets and the forwarding of others' tweets (known as retweets). Past work on examining Twitter followers has looked at the distribution of various network attributes (i.e., in and outward degrees, clustering coefficients, and betweenness) among a sample of users. The analysis of these metrics show that the vast majority of users keep a similar number of followers as those that they follow, but that nearly half of Twitter accounts are poorly connected to the network as a whole — considered to be leaf nodes who are inactive.[2] When we consider how Twitter and other social networks are related to in-person human interactions, looking at how users follow each other can be deceiving because it is so easy for users to begin following a set of users but then only read and follow up on the tweets by a subset of these users. Thus, the act of simply following users is passive and not necessarily representative of human relationships; it is the same as if, say on Facebook, one friends one thousand users — it is unlikely that this person converses with all thousand on a regular basis.

This is why an analysis of the retweet and mentions network is crucial to understanding the human dynamics present within Twitter. The path of retweet propagation can be traced and compared to typical human networks. After all, retweeting a message is much the same as retelling some words to a friend. As Bild et al have mentioned, "the retweet graph may better encode true interest and

[1] Gephi: Datasets, <https://wiki.gephi.org/index.php/Datasets>.

[2] Abraham Ronel Martínez Teutle, "Twitter: Network Properties Analysis," Universidad de las Américas Puebla, <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05440773>.

trust relationships among users."[3] Work on retweet networks over the past several years has yielded findings on what sort of content is more likely to be forwarded on by other users (typically content with embedded URLs or clickable keywords — hashtags, as they are called). Further, the retweet network has been characterized by Bild et al as having various small-world characteristics, with several approaching the typical ranges of scale-free networks, which are themselves considered to be ultra-small-world networks. Compared with the network of all Twitter followers and followings, the retweet graph is said to be "less disassortative and more highly clustered." These are the sort of findings that we will attempt to replicate and comment further on in this exercise.

Methods and results

In order to conduct our analysis, we chose to work with Gephi-provided "Twitter Mentions and Retweets" network. This is a network containing 3658 nodes and 188712 edges, displayed in a format showing, for each entry, a Node A string, a Node B string, and an edge "weight" value. Each node represent a Twitter user, and each edge denotes how frequently a user ("Source A") mentions or retweets from another user ("Target B") as defined by the weight property; therefore, we face a directed graph of different users.

Using the NetworkX package for Python, we measure several properties of the retweet and mentions network, generated degree distribution plot, and compared our network with several typical models. We also attempted to find the relationship between degree and betweenness centrality, but, as to be shown, this result was not very clear-cut.

Degree distribution

We first calculated the total degree (both inward and outward) for each of 3658 nodes and plotted them as a distribution on a simple linear scale. As we see in Figure 1, the shape resembles that of a power law distribution (which, if found, would be a tip-off to a scale-free network):

[3] David R Bild et al, "Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph," University of Michigan, Stellar@MIT.

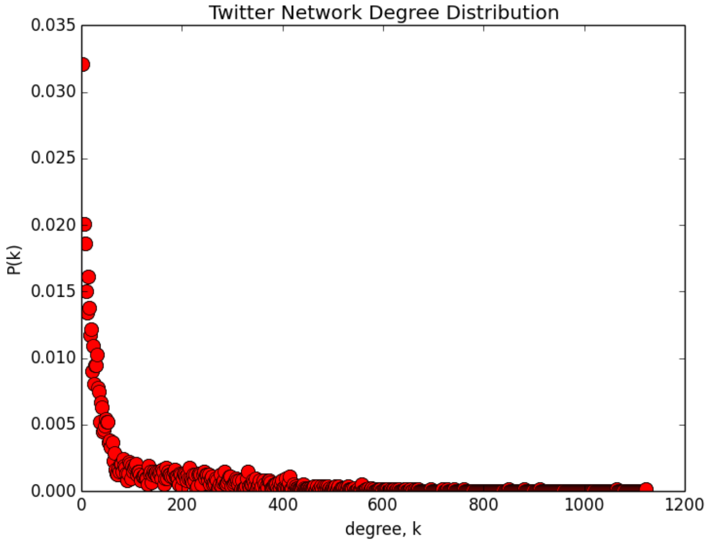

Figure 1, linear degree distribution of retweet and mentions network.

To test the theory of a power law distribution of nodal degrees, we fit this same data onto a graph of logarithmic scale, and binned the data points so as to break up the graph into a grid and take averages at regular intervals so as to approximate single function to the data. The result is that the network data fits well to a power distribution of constant scaling factor -2.0 and gamma exponential factor of -1.1 (as shown by the dashed line), at least until the users' degree exceeds roughly 600. This appears to be the cut-off for the majority of Twitter users — those with more than 600 people retweeting them are in the vast minority, although (when taking into account the full network) they make up many well-known personalities and news organizations. Regardless, our logarithmic fitting suggests that the scale-free model might be a suitable description of the network.

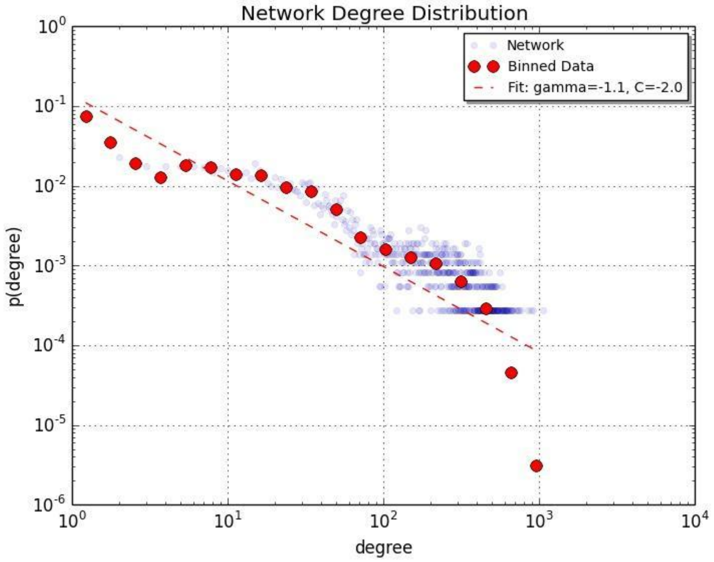

Figure 2, binned logarithmic distribution of nodal degrees.

Betweenness versus degree

We tried to gain a deeper insight by taking a look at the relationship between the betweenness centrality and degree for every node in this network. We conjectured that nodes with higher degrees also have higher values of betweenness, as they are more likely to lie on the shortest paths between more pairs of users. Using Python, we were able to plot the diagram (figure 3). The plot shows the following:

- Points with low betweenness (<0.005) have degrees ranging from as low as 1 to as high as 900
- Points with low degrees (<200) might have values of betweenness ranging from 0 to as high as 0.015
- Few points have both high degrees and high values of betweenness, but most points have low betweenness regardless of their degrees

It is thus difficult to tell whether a relationship exists between the two quantities. We might be able to discover the relationship and verify our conjecture with a bigger dataset.

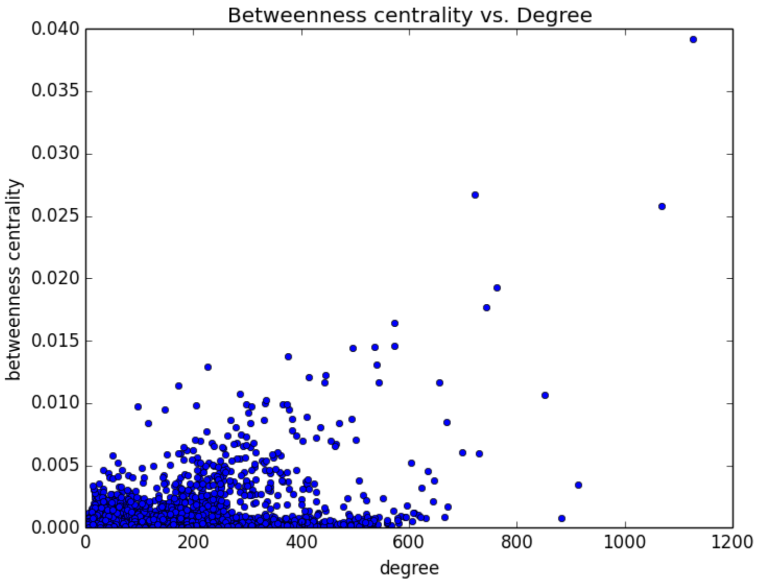

Figure 3, betweenness centrality of each node plot against its respective degree.

Other network attributes

Using Gephi and Python we also took stock of several other common network attributes:

Average nodal degree	103.234
Average clustering coefficient	0.319
Average shortest path length	3.764

Table 1, Twitter retweet network attributes.

These numbers put us in line with others' past findings on Twitter networks, showing the average clustering coefficient to be around 0.3.[4] Other online social networks are also in similar ranges, such as Flickr (a web service with a premise on sharing personal photographs) at 0.40.[5] Finally, data on the average path length between Twitter users (discussed further under Conclusions) show that our findings are on the correct order of magnitude, which helps to verify that we have a valid sampling of the Twitter network.

Comparison with network models

We compared our graph with several typical network models:

	Twittercrawl	Small World	Random Graph	Barabási-Albert
Average degree	103.234	103.234	$< k > = pN = 103.234, p = 0.02822$	103.178
Average shortest path length	3.764	$\sim \frac{\log N}{\log K} = 1.769$	$\sim \frac{\log N}{\log K} = 1.769$	~6.46
Average clustering coefficient	0.319	$0.743(1-p)^3 = 0.319, p = 0.246$	$< CC >= p = 0.02822$	~0.018
Degree distribution	Power law	Delta to Poissonian centered around 103.234	Binomial to Poissonian centered around 103.234	Power law

Table 2, comparison of retweet graph to common network models.

[4] John Schwartz, "Properties of Twitter," Rensselaer Polytechnic Institute, <http://www.cs.rpi.edu/~magdon/courses/casp/projects/Schwarz.pdf>.

[5] Jérôme Kunegis, "Clustering coefficient," Universität Koblenz-Landau, <http://konect.uni-koblenz.de/statistics/clusco>.

The scale-free model is the most descriptive one for our network. It predicts the value of average degree accurately and has a degree distribution similar to that of our network. The two other network models would predict that most of the nodes have degrees close to the average degree 103.234, i.e. every node tends to behave like an average. However, it is the nature of Twitter that most of its users are ordinary people with limited power of influence; therefore the scale-free model best captures such characteristics.

Finally, we wished to visualize the retweet graph in a manner that would highlight the big picture of principle nodes. This involved the deletion of nodes with small weights in order to cut out users that are not active contributors or who had abandoned their accounts shortly after their creation (and thus may skew the graph). A force-directed graph algorithm (ForceAtlas in Gephi) was applied to better organize the network and its relationships, with any lingering unconnected nodes deleted in order to show only active users. The final result is shown below as Figure 4:

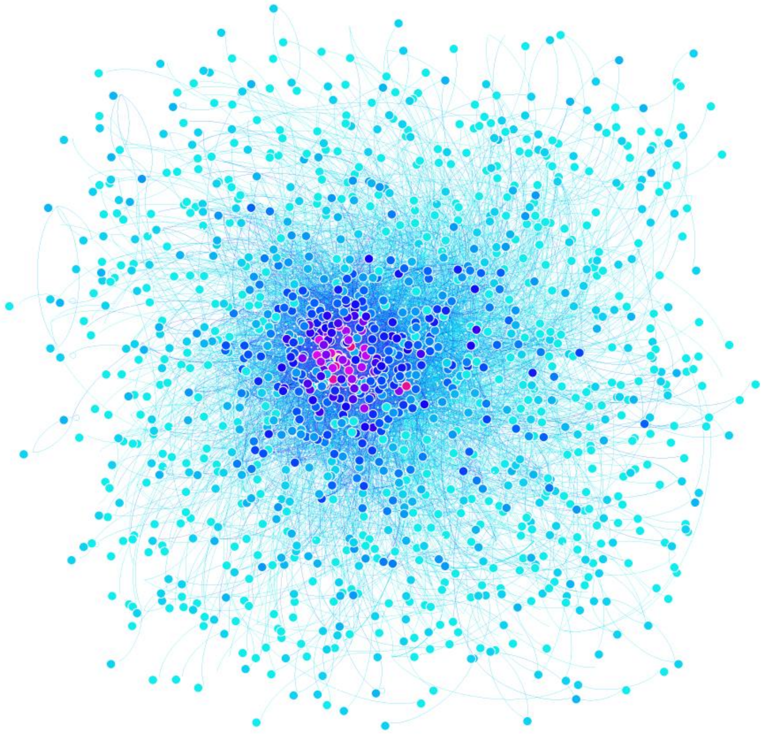

Figure 4, visualization of retweet network excepting nodes with no connectivity and links with very small weight.

The visualization shows that many of the nodes with greater degrees (darker dots) also happen to be linked by edges with higher weights (darker links). It follows our intuition that popular accounts are more likely to be information hubs, i.e. whose tweets are spread by more users.

Conclusions

Our network has both scale-free and small-world properties.

The degree distribution and the nature of our network suggest the scale-free property. As scale-free networks are generated by growth and preferential attachment, so is our network. The number of mentions and retweets increases with time (growth), and users are more likely to retweet from popular accounts (preferential attachment).

On the other hand, our network is also small-world. The average shortest path of the network is 3.764, which is a four-degree separation. This is only slightly lower than the average shortest path of the entire Twitter followers network of 4.1. While we expect the retweet network to be well-connected, the fact that these numbers are similar despite the differences in networks analyzed suggest that a larger sampling size may yield an even smaller path length (since these interactions tend to be more correlated with human relationships.[6]

Applications and future studies

The primary limitation in this report was the small size of the Twitter dataset we had access to, representing 0.001% of all active users. Other studies, such as the one by Kwak et al, crawled over 40 million users in 2009, although such compilation of data takes many days due to bandwidth limits put in place by Twitter. Nevertheless, we have found that the average shortest path length and other network properties of our set is not too dissimilar from that obtained in other research, thus verifying that our obtained sample was at least relatively random and unbiased given its small size.

This dataset can also reveal other interactions with other Twitter datasets. For example, we can find out to what extent Twitter users consider people they follow as reliable sources of information by comparing the Twitter following network with the retweet/mention network. Possible analyses include:

- Comparing the two networks and finding out how properties of one network can be applied to the other
- Investigating why Twitter accounts with similar numbers of followers and contents have different mentions/retweet values
- Exploring how messages spread in Twitter, given the content of each retweet/mention

There are, of course, other applications of determining the nature of the retweet network. For instance, Twitter, as a web service, can benefit from the ability to predict the behaviors of users and

[6] Haewood Kwak et al, “What is Twitter, a Social Network or a News Media?,” 30 April 2010, <http://an.kaist.ac.kr/traces/WWW2010.html>.

thus more-easily detect spam accounts. These accounts, often computer-generated, are much less connected than even a new user that has several human counterparts on the network.

Linking geography to the virtual network is another area in which further study may generate new insights. The Changing Places group at the MIT Media Lab, has prototyped an interactive and physical model of Kendall Square, onto which different information can be projected, including real-time tweets that are geo-tagged to the neighborhood. It may conceived, knowing typical network parameters, that determining the connectedness of the retweets in a geographic area can provide an indication of that location's social functionality and diversity.