The Analysis of a Retweet Network

Background

On the 25th of February 2013, Donald Trump, the famous Real Estate tycoon, tweeted a comment about the famous TV-show 'Shark Tank'. One of his followers, Mark copied the link to Donald tweet and added a prefix 'Wow'. One of Mark's followers, Erik, then copied the link to this tweet by Mark, also adding the prefix 'Wow'. Just like that, this retweeting game took off and more people got involved in this 'retweet chain'. Fig. 1 shows how the chain began, from Donald to Mark to Erik. I carried out an analysis on this retweet network, and this is a report detailing my findings.

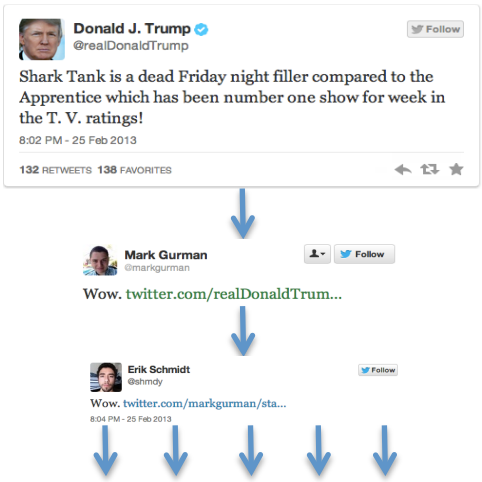

Fig 1: The beginning of the retweet chain

Analysis

The network data of this retweet network was obtained by a twitter crawl. Nodes in this network represent the twitter users, while the edges represent the retweets. The network is a directional one, with the directionality such that Erik links to Mark as Erik did the retweeting. The Twitter network has 651 nodes (users) and 634 edges (retweets). Fig 2 is a visualization of this network. The colors of the nodes are determined by their degree, on a scale from blue (low degree) to red (high degree). I analyzed this network in comparison to a scale free network (shown in Fig 3) with

the same number of nodes, but 1300 edges. The colors of the nodes are also categorized in a similar way, on the blue – red scale, depending on their degree.

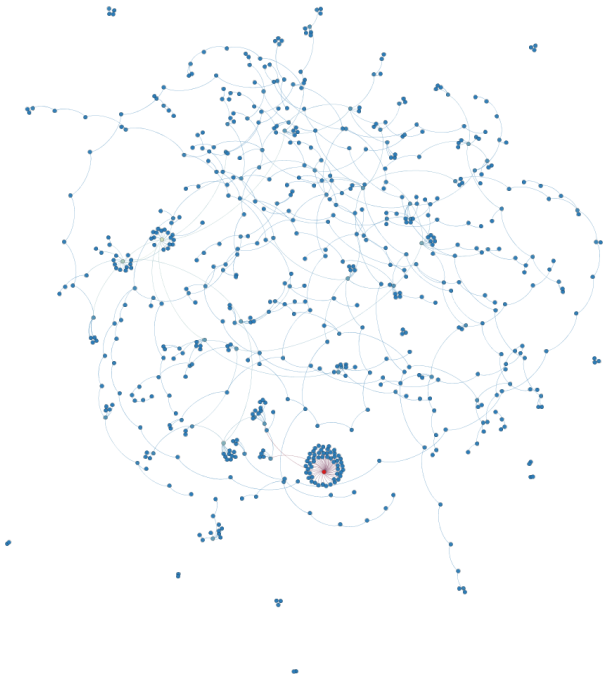

Fig 2: Visualization of the twitter network

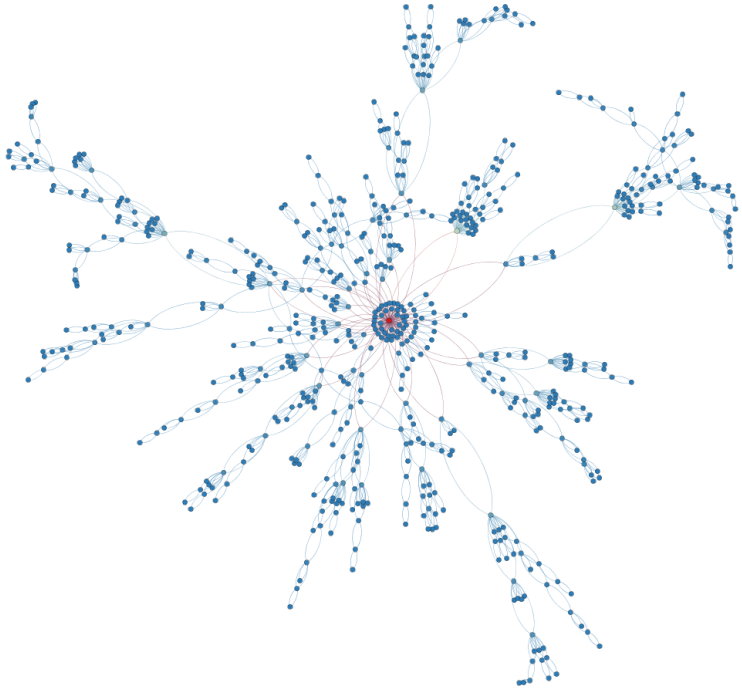

Fig 2: Visualization of the Scale Free

Straight away we can draw contrasts between the twitter network and the scale free network. We see that the twitter network is made up mostly of long single node chains, where one node is connected to just one other node; this is consistent with the average degree of a node being about 0.974. On the other hand, in the Scale Free network, on average, the nodes have an average degree of 1.997. It can be seen from the visualization of the scale free network that this 'long chain' property doesn't exactly exist in the scale free network.

Using gephi, I calculated the average path length of both networks; the twitter network had an average path length of 14.428. This means that, on average, one user in the twitter network is about 14 retweets away from another user. The scale free network, on the other hand, has average path length of 6.236. This difference between attributable to the low average degree that the twitter network has, with most of the users being retweeted by just one person. Fig 4 is a table showing some important network properties of both of the networks.

	Number of Nodes	Number of Edges	Average Degree	Average Path Length	Network Diameter	Degree Distribution
Twitter Network	651	634	0.974	14.428	55	Power Law
Scale-Free Network	651	1300	1.997	6.236	15	Power Law

Fig 4: Table showing important network properties

In addition to these calculations, Fig 5 shows the degree distribution of both the scale-free network and the twitter network. The relationship in both cases is clearly follows the power law behavior. This is a characteristic of the scale free network, so this is what I anticipated in this case. For the twitter network however, this relationship tells us that users with fewer retweets are more prevalent than users with more retweets. This can be seen in the visualization where there are many blue

nodes but only one red node. The twitter network thus resembles the scale free network rather closely in this regard.

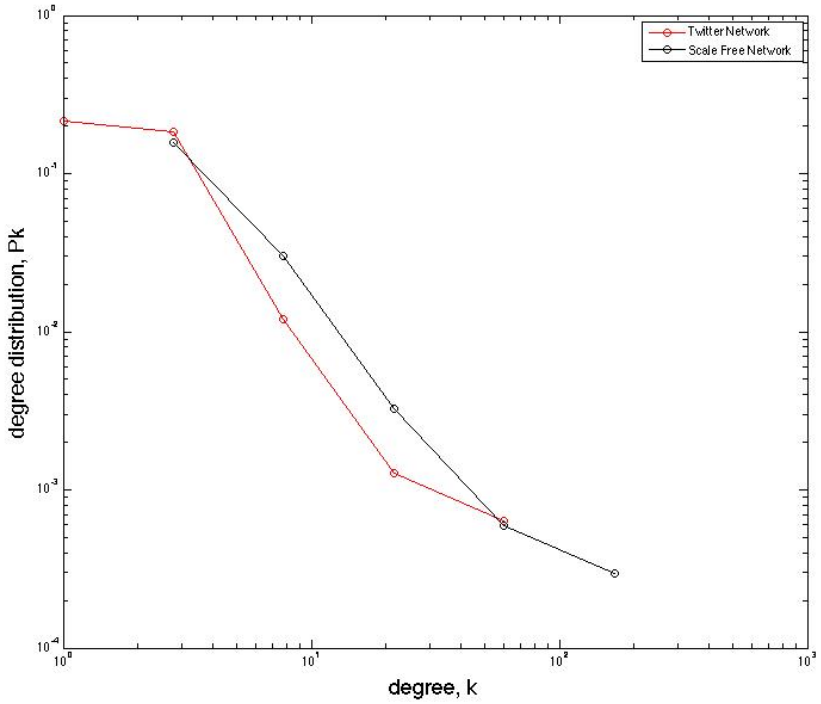

Fig 5: Degree distribution of twitter network and scale free network

Furthermore, I explored the relationships between the degrees of nodes and their degree of nearest neighbors. Fig 6 shows this relationship. Both plots are similar. They have a larger number of data points at the beginning varying from low to high Knn's, but have smaller number of points as the degree increases. This can be explained by the degree distribution plot in Fig 5, which showed that there are more nodes with low degrees than high degrees. Fig 7 and Fig 8 show the Knn distribution in a log-log plot for the twitter network and the scale free network respectively. There is a pseudo trend in the average Knn; being high for low degrees and low for high degrees; this can be seen in both the twitter network and the scale free network alike. In the case of the twitter network, this tells us that people with fewer retweets are, on average, retweeting people with more retweets, and people with more retweets are being retweeted by people with fewer retweets. For example, Donald Trump is being retweeted by a lot of people; some with a number of

retweets, but most with no further retweets, so his Knn will be very close to one. Another user, for example who only retweets Trump, and has no further retweets has only Donald as his nearest neighbor, meaning his Knn will be the number of retweets that Donald has; which in this case is about 55. This tells us that in compliance with the Barabasi-Albert preferential attachment, more users retweeting people with more followers than those with less. 'The rich get richer, and the poor don't get rich."

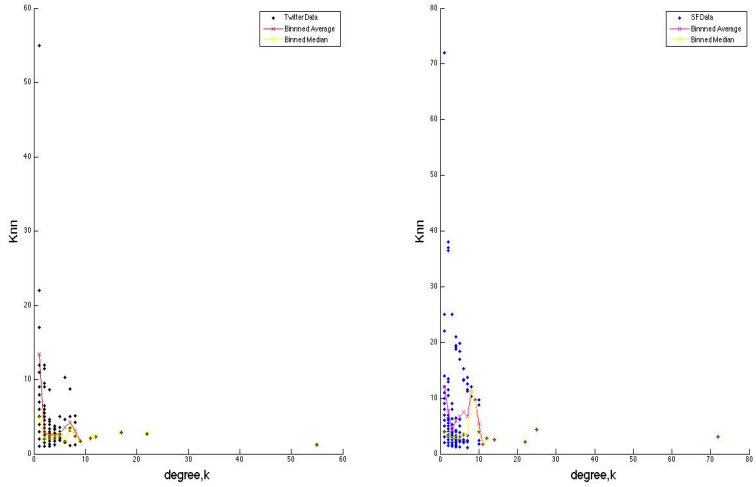

Fig 6: The relationship between the degree and the degree of nearest neighbor of a node

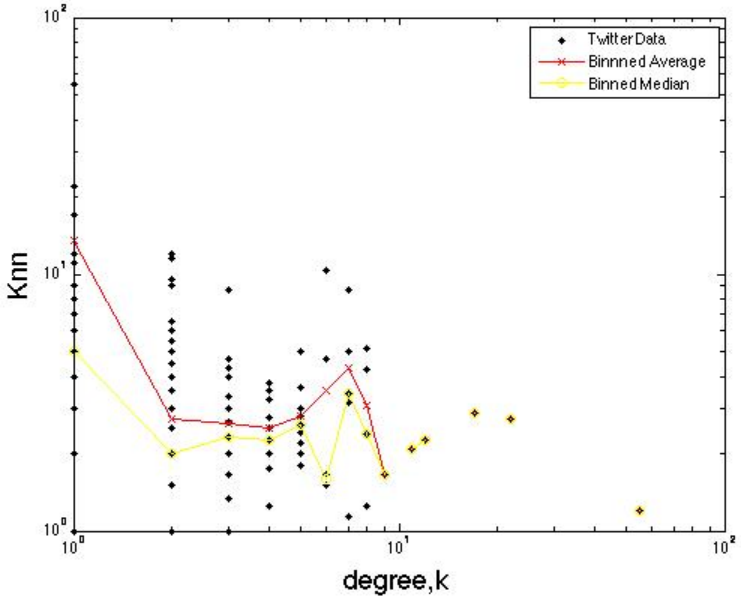

Fig 7: Log-Log plot showing the relationship between the degree and the degree of nearest neighbor of a node in the twitter network

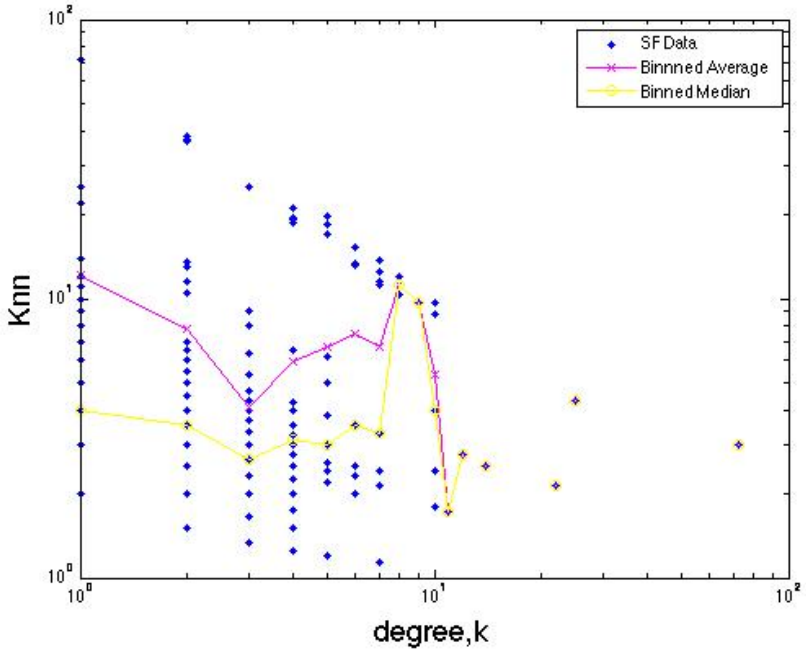

Fig 7: Log-Log plot showing the relationship between the degree and the degree of nearest neighbor of a node in the scale-free network

Related Work and Further Research

During my research, I came across a paper called 'Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number" by Bruno Goncalves, Nicola Perra and Alesandro Vespignani. Their paper sought to show that the maximum number of stable interpersonal cyber relationships that a person can maintain is a maximum within the range of 100 to 200 people. I propose future research that shows the relationship between activity on social media and the number of stable relationships a person is able to maintain. Are people who maintain more relationships generally more active than others? I believe this will be interesting to investigate further.