The Network Analysis of Meet-Up Members in Nashville

DNSC 6215

Social Network Analytics

Mengqi Li, Xuan Yang

Feb 28th, 2018

**Abstract**

This project focuses on the social network relationships between users on meetup.com located in Nashville. Multiple algorithms and evaluation methods like graph-based social network analysis and centrality measures were exercised to fulfill the purpose of disclosing the reality in the social network media. Rely on the graph and diversified measures, the differing in the distribution of social network resources were demonstrated as a “social gap” between distinct kind of participators.

**Introduction**

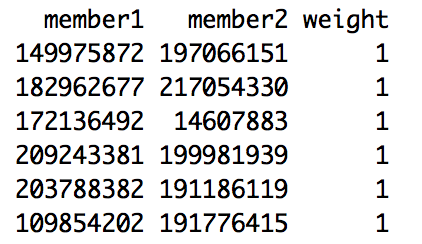
Today, social media like Facebook and Twitter has been possessing a growing number of our time for people of all ages. According to an online investigation, the daily usage of social media grows from 90 minutes in 2012 to 135 minutes in 2017 which is quite a lot of time per day (“Daily-social-media-usage-worldwide”, 2017). Considering the spreading influence of the social media, loads of researches have been conducted to ascertain some regular patterns behind the glamorous social media phenomenon.

It is indisputable that there are amounts of opportunities behind the infinite resources, people around the world are currently digging data to serve various kinds of purposes. As students in social network analytics course, we are more than curious to analyze our own topic which is the meetup.com website data located in Nashville. The Meetup.com is a website for people to organize and participate in assorted social activities regularly. In that case, the relationships among all its users and the people they meet during all the activities could be a perfect sample for us to do graph-based social network analysis. And that is exactly what we have learned in this course and we did the following analysis to extract insights from the meetup.com in Nashville and in the meantime unveiling the social network reality in the cyberspace.

**Data Overview**

The dataset used in the project comes from Kaggle (Bailey). There are two csv in the dataset. One is members’ edges, and the other is metadata about members. The original csv has more than 20,000 nodes. In order to avoid insignificant results, we randomly selected 1,000 nodes and associated edges. In addition, we decomposed the graph plotted based on the sample data, choosing the most typical component of the network to perform further analysis.

**Figure 1**



Here is a table displaying the sample dataset where each node is representing a single member. Each member has his or her own id number, and each edge in between them indicated whether member 1 knew member 2. The weight in the rightmost column stands for the exact number of events that both member 1 and member 2 attended in the history.

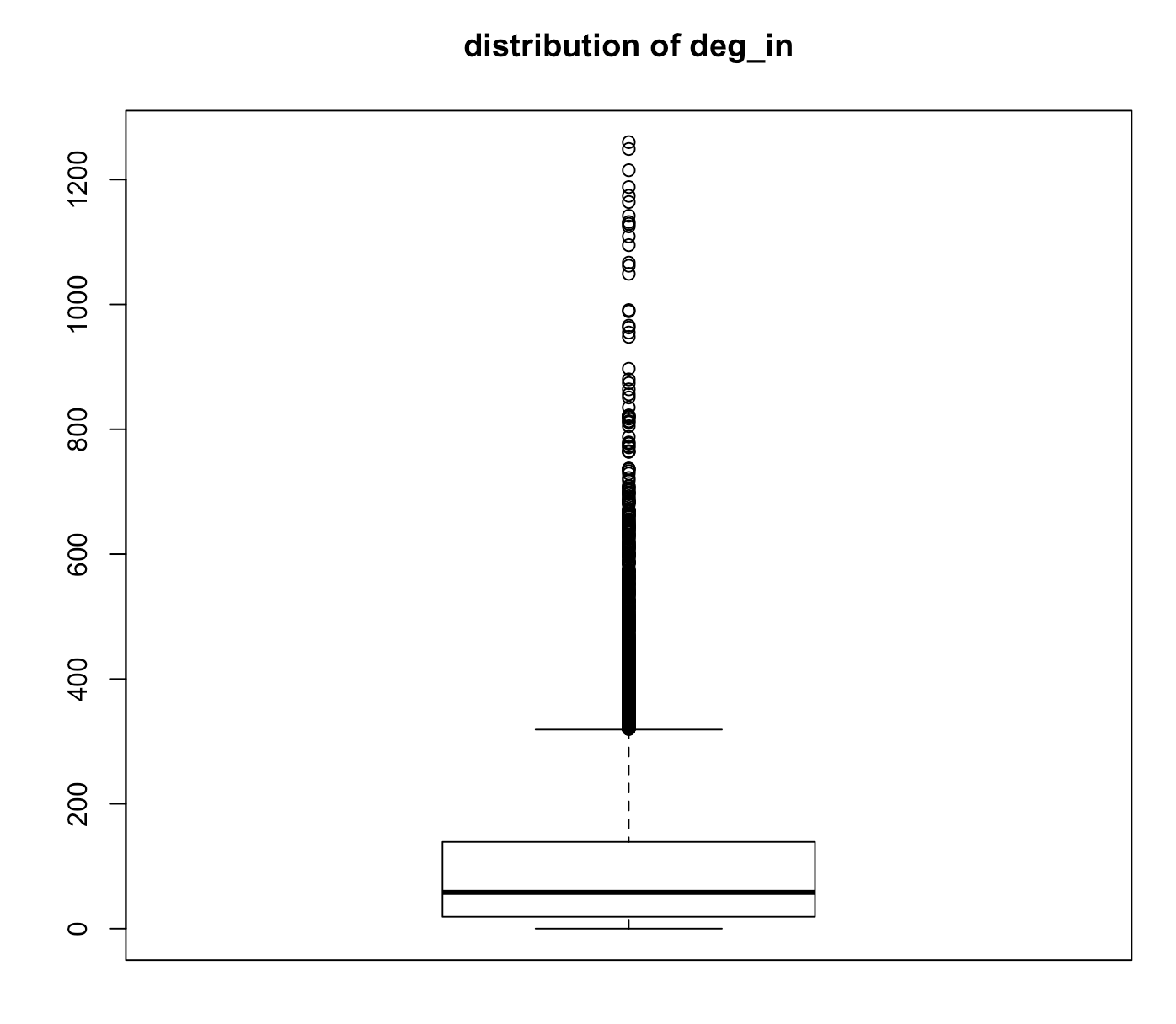
**Algorithms and Models**

The algorithms and models used in the social media project incorporated a social network graph depending on the chosen sample dataset to present a basic understanding of the relationships we are interested in, and in-degree and out-degree centrality as data estimators to reveal the number of node which means website users in reality related to each vertex with regarding to direction, closeness and betweenness centrality were adopted to determine the popularity in consider of the length and number of the link between each node, eigenvector centrality was utilized to measure the social importance of a specific node in the social network. Correlation plot was also applied in order to find out the dependence between all the estimations like our centrality measures above. Furthermore, other evaluation methods like shortest path, reachability, reciprocity, transitivity, and coreness were also exercised to identify the facts and patterns in the social media data.

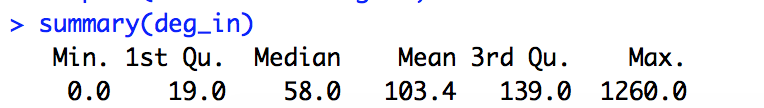
**Network Analysis**

To be specific, the following output is evidence which derives our final conclusions. At first, when we look at the centrality measurement. According to boxplots of in-degree (figure 2) and out-degree distribution (figure 4) below, we found no matter in or out, there are lots of people have zero degrees. In the in-degree plot, although there are many outliers, most people have less than 400 degrees, and according to data summary (figure 3), the mean value is 103.4. According to output about out-degree centrality, the distribution is similar to in-degree centrality. There are also lots of people have zero out-degree and the maximum is 1987. The mean is 103.4. Although the 3rd quartile is 138 there are lots of outliers. The interesting result we found that the person who has the most in-degree just has the average level value of out-degree. The same thing happens to out-degree.

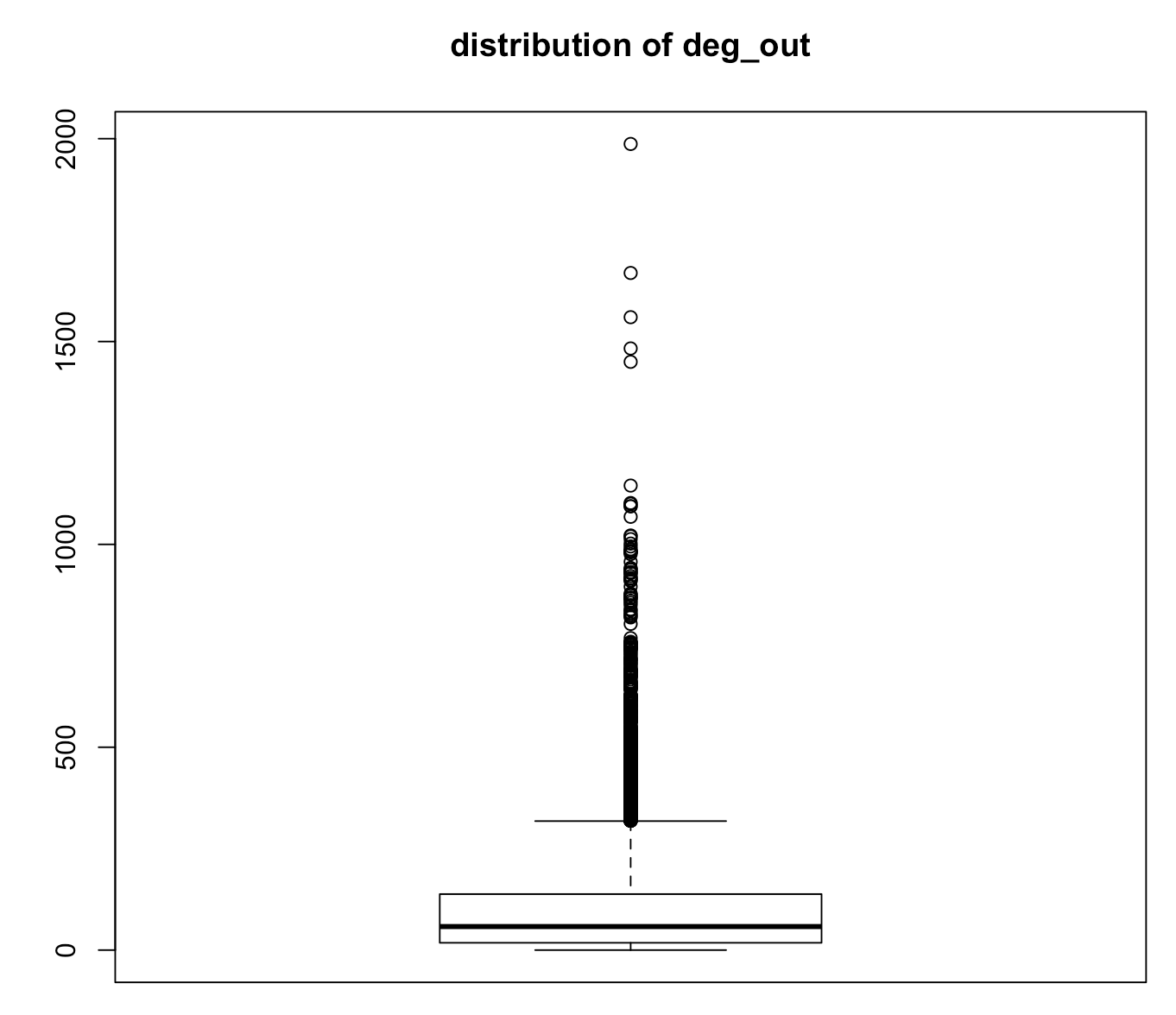
**Figure 2**



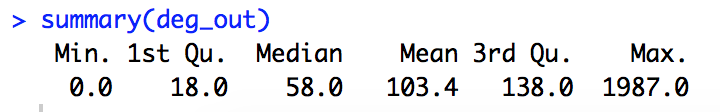
**Figure 3**



**Figure 4**

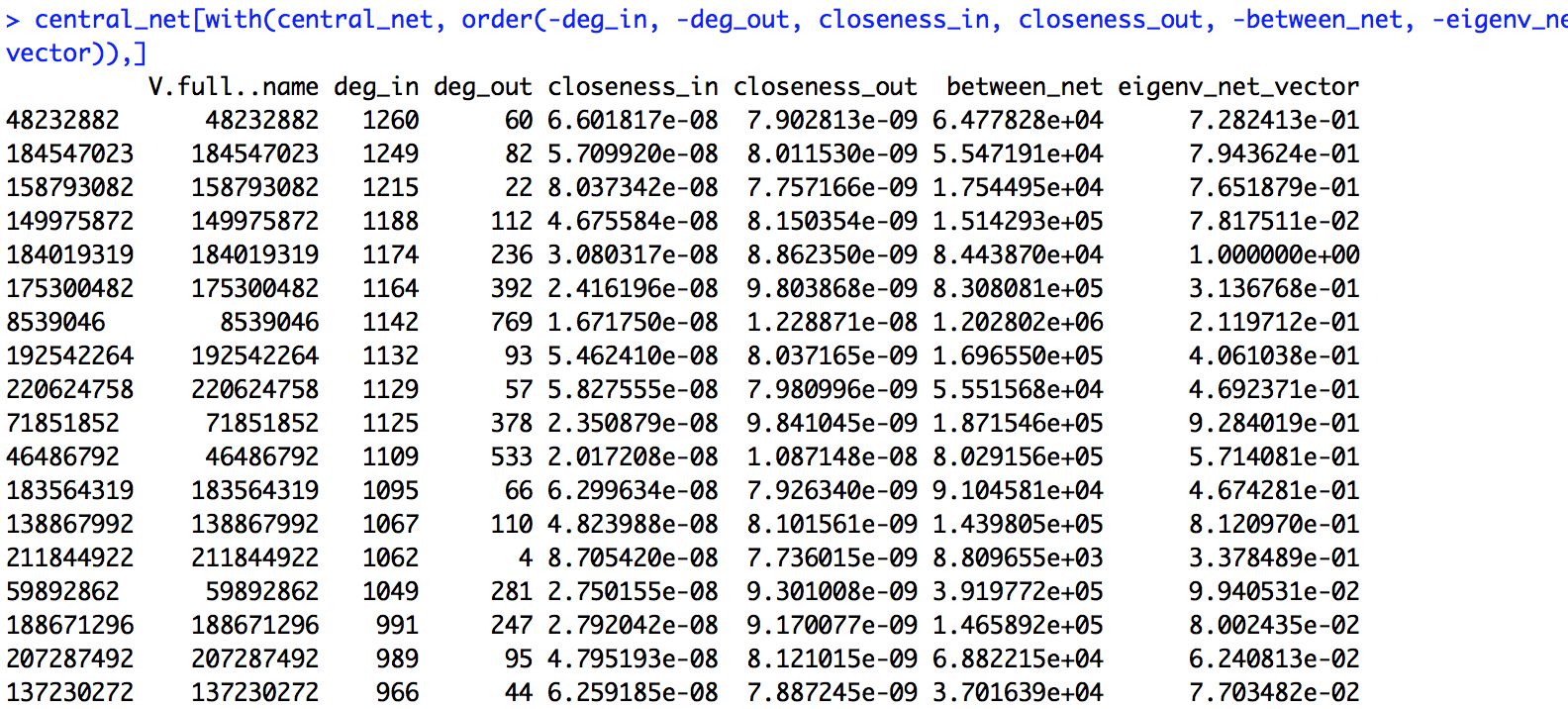


**Figure 5**



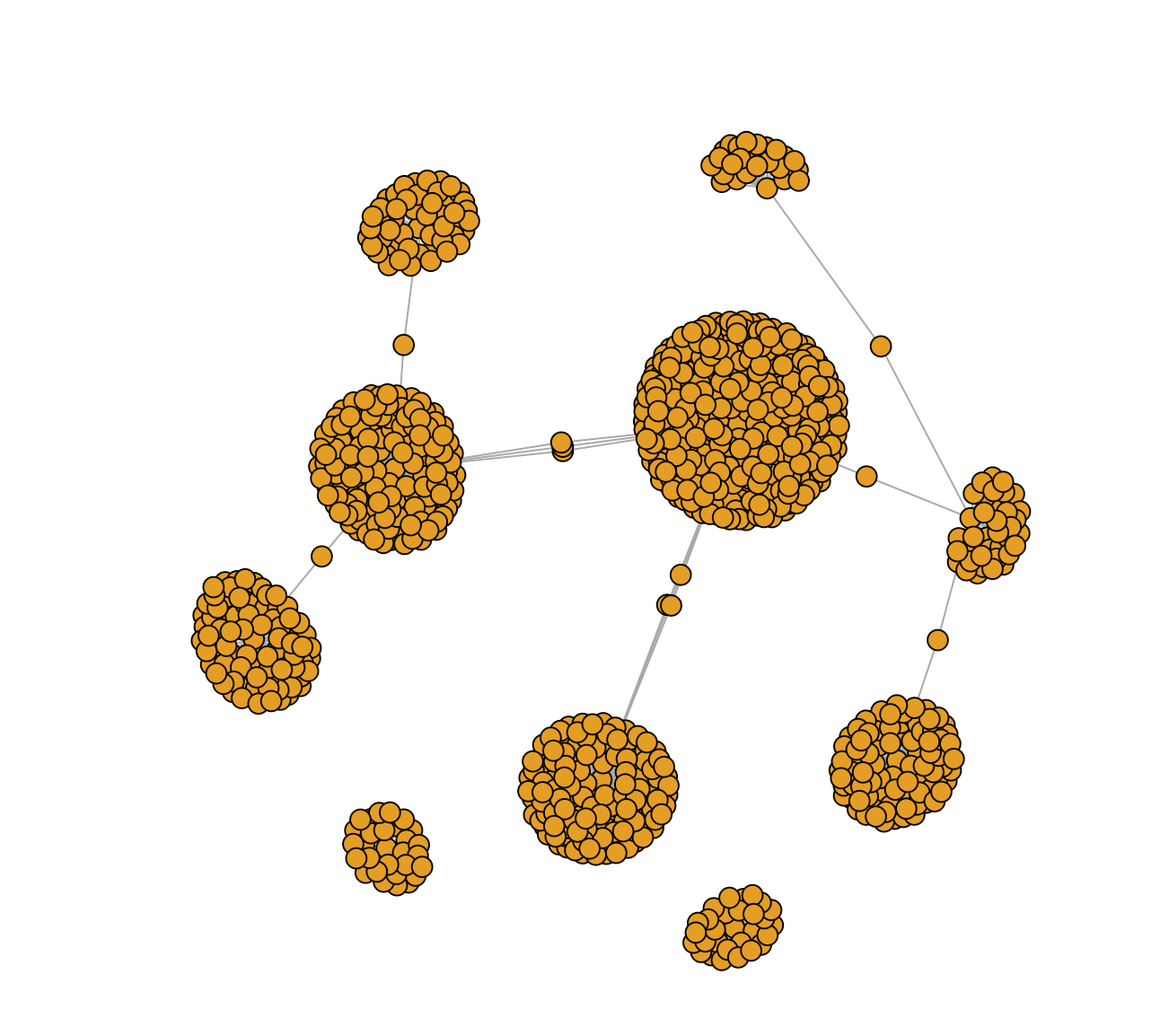
When we order the data by all of the centrality parameters, like the highest degree, the smallest closeness, the highest betweenness and highest eigenvector. We found the person (shown in figure 6) who could be regarded as the most popular member of the Meet-Up. And this person has very tiny closeness and more than 6000 shortest paths. Besides, the person has 0.7 eigenvector weight.

**Figure 6**



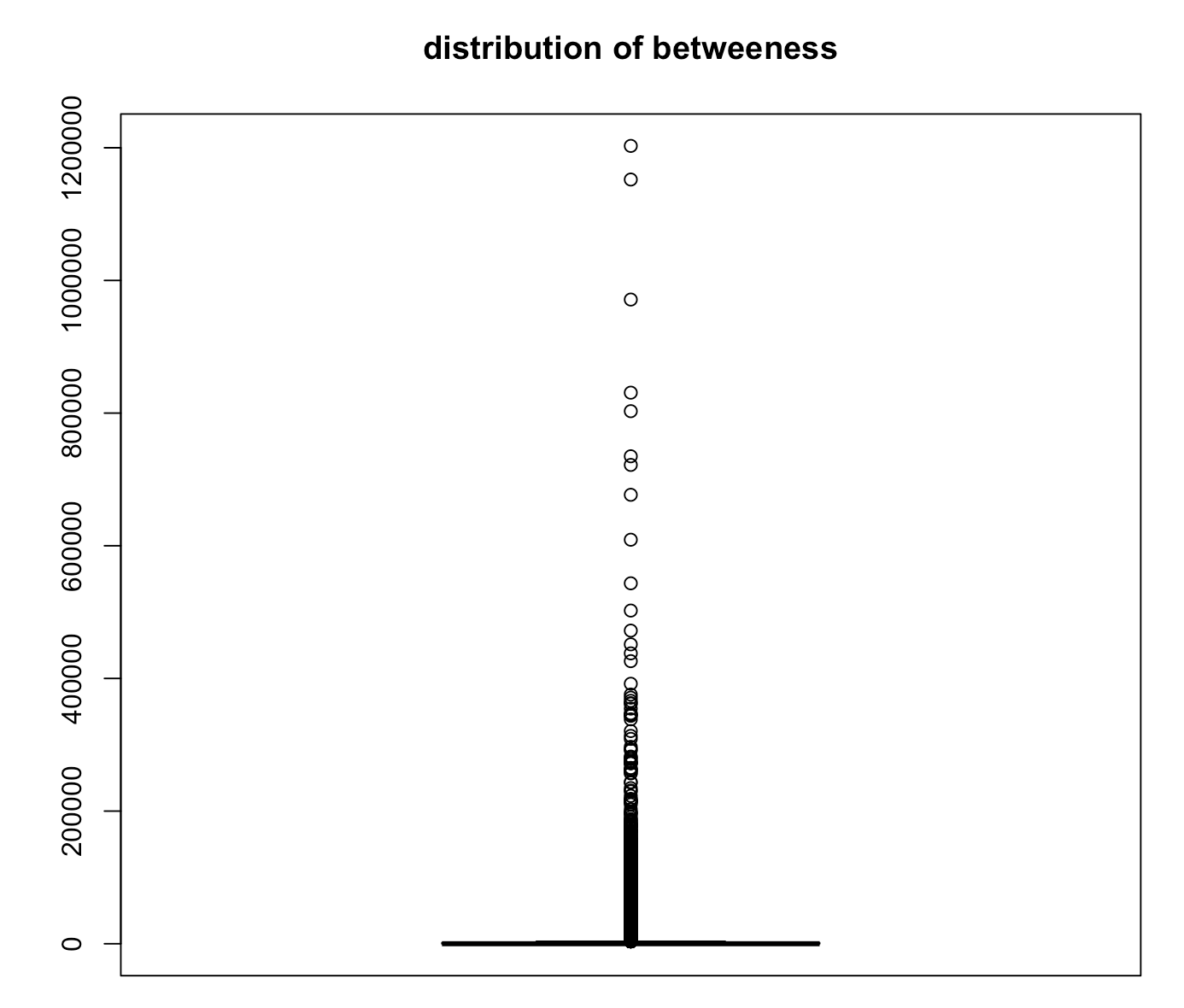
According to the network graphs (shown in figure 7) we plotted, we found in this dataset, there are many separated clusters. They are connected by some individuals or just isolated. Thus, we take the next step to look deeper into betweenness centrality.

**Figure 7**

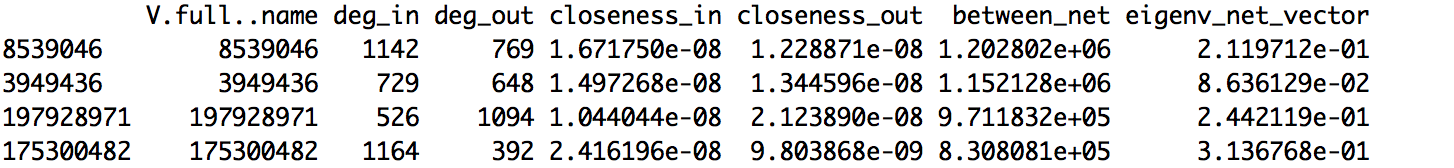


When we calculate betweenness centrality, we found the betweenness also distributed very unevenly (shown in figure 8 and figure 9). The highest node of betweenness is more than 1 million. But the lowest node of betweenness is just zero. And when we merge the edge file with meta member information file, we found there are 3 people are from Nashville among the top 4 betweenness guys (shown in figure 10). Thus, we assume that there is a relationship between centrality and city parameter and Meet-Up is a social networking website which is very regional popular. There are might be a group of active members come from a certain city commonly and they become the community bridge linking between different clusters.

**Figure 8**



**Figure 9**

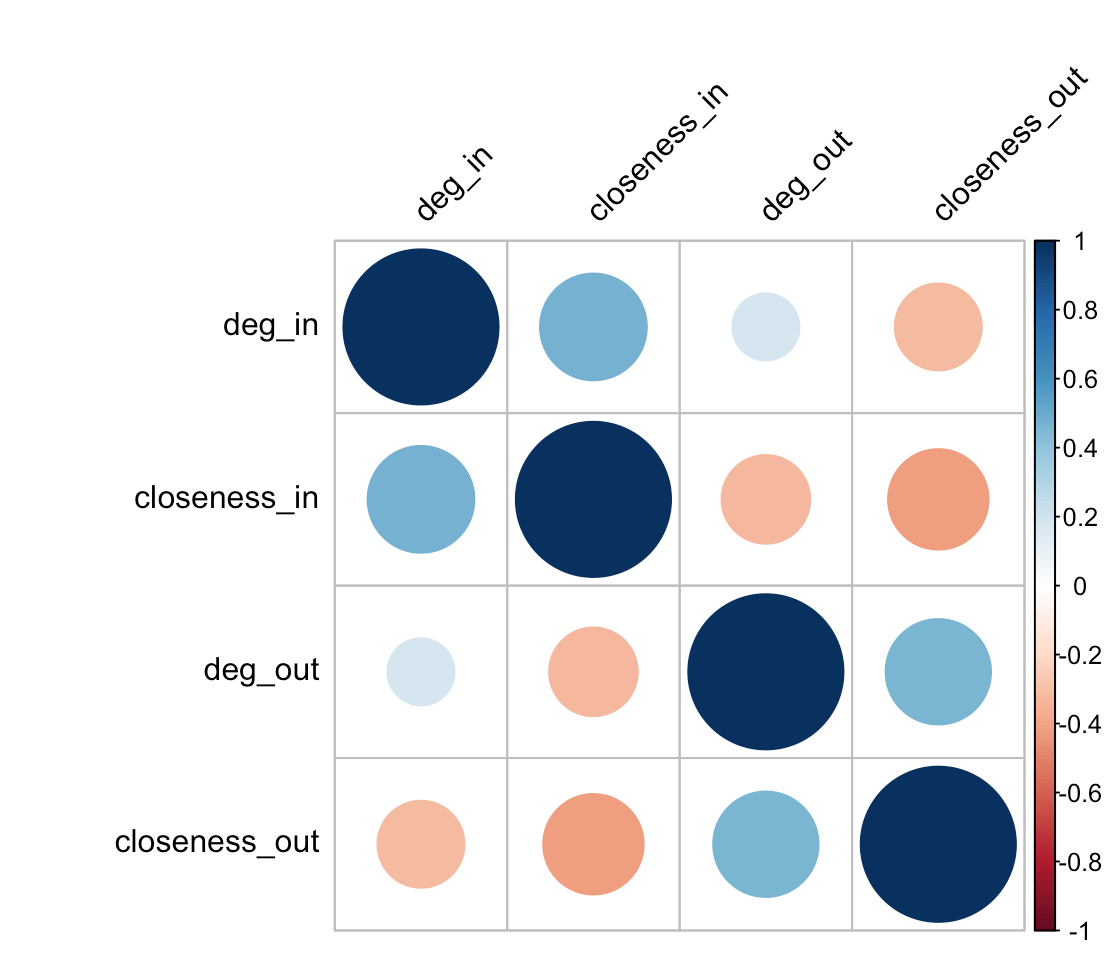


**Figure 10**

|  |  |  |
| --- | --- | --- |
| ID | First Name | City |
| 8539046 | Shalini | Nashville |
| 3949436 | Pablo | Nashville |
| 197928971 | Ted | Franklin |
| 175300482 | Rav | Nashville |

In the next step, we implement correlation analysis among centrality. As is shown in figure 11 here, it tells all correlations between each pair of centrality parameters. Specifically, closeness in and out has a negative relation, degree out and closeness in, plus degree in and closeness out have a negative relationship. Others are positive. But there is no significant difference between those correlations. This correlation result displays beyond our expectation, and the reasons behind these still need our further exploration.

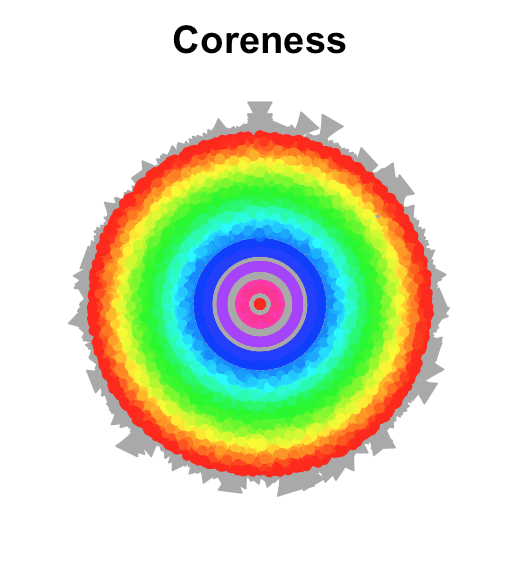
**Figure 11**



To get a more accurate statistics index, we decompose the whole dataset and pick up the largest one. The following are calculated result: the reachability in the social network example here, the mean of the full model in regard to whether in or out is 0.002003004 which means for most of the node in the graph, they are only linked to a few amount of the other node.

By calculating the coreness, we plot the result (shown in figure 12) and find that there does exist a network center and they are tightly interlinked within the network.

**Figure 12**



**Challenges and Solutions**

The challenges in completing this project are mainly caused by the large sample size of the data, as it contains 1,000 nodes and 50,000 edges even after the random selections. The result of many estimates was not satisfying as it did not disclose any particular patterns and it took quite a long time to produce a reasonable explanation for us to come up with a clear and consistent conclusion for the whole project. In contemplation of a compelling consequence, the largest component in the social network example, as well as the most symbolic one, was applied to our interpretations.

**Conclusion**

Generally speaking, the main findings of our social network analysis is that there do exist a “social gap” between peoples like the crucial “income gap” in the real world which simply suggested that most people are not closely tied with many people out of their reach or even isolated on a relative basis whereas there also remained quite a few of them are highly popular and occupy most of the social network resources based on the fact that they have high centrality and multiple edges between numerous people. Moreover, the clusters in the graph might be influenced by locations people live in. By the same token, the significance of the people who perform as bridges linking between clusters is considerably evident as they were also the ones who retained most of the social network resources.

**References**

1. Bailey, Stephen. “Nashville Meetup Network.” *Kaggle.* 2018.
2. Daily time spent on social networking by internet users worldwide from 2012 to 2017 (in minutes). *The Statistics Portal.* 2017.
3. McFarland, Daniel, Solomon Messing, Michael Nowak, and Sean J. Westwood. “Social Network Analysis Labs in R.” *Stanford University.* 2010.
4. Vladimir Batagelj, Matjaz Zaversnik. “An O(m) Algorithm for Cores Decomposition of Networks.” *Recent Trends in Graph Theory, Algebraic Combinatorics, and Graph Algorithms*. 2002.