July 12, 2016

Yuhui Zhang

**Twitter pre-class assignment**

### *Network structure visualization*

The output of the network structure of the subset edges is shown below, to the left is a simply plot of network topology and to the right is a plot with color coded communities calculated with Walktrap community finding algorithm.

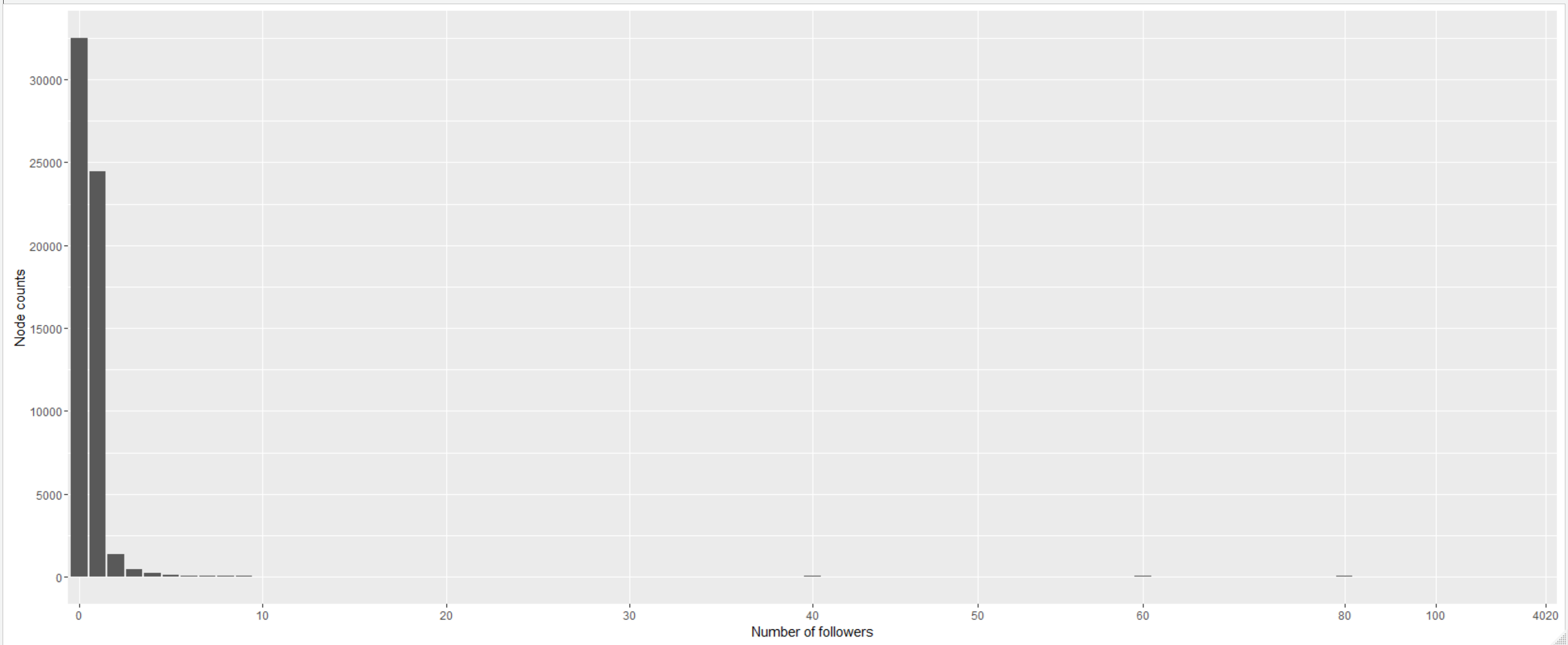
|  |  |
| --- | --- |
|  |  |

We can see from the network structure that most of the users subscribes to four vertexes (0, 1, 2, and 3) and there are many much smaller components formed by vertex connecting with various directions

### *Data Analysis*

###### 1. The distribution of the of number of followers

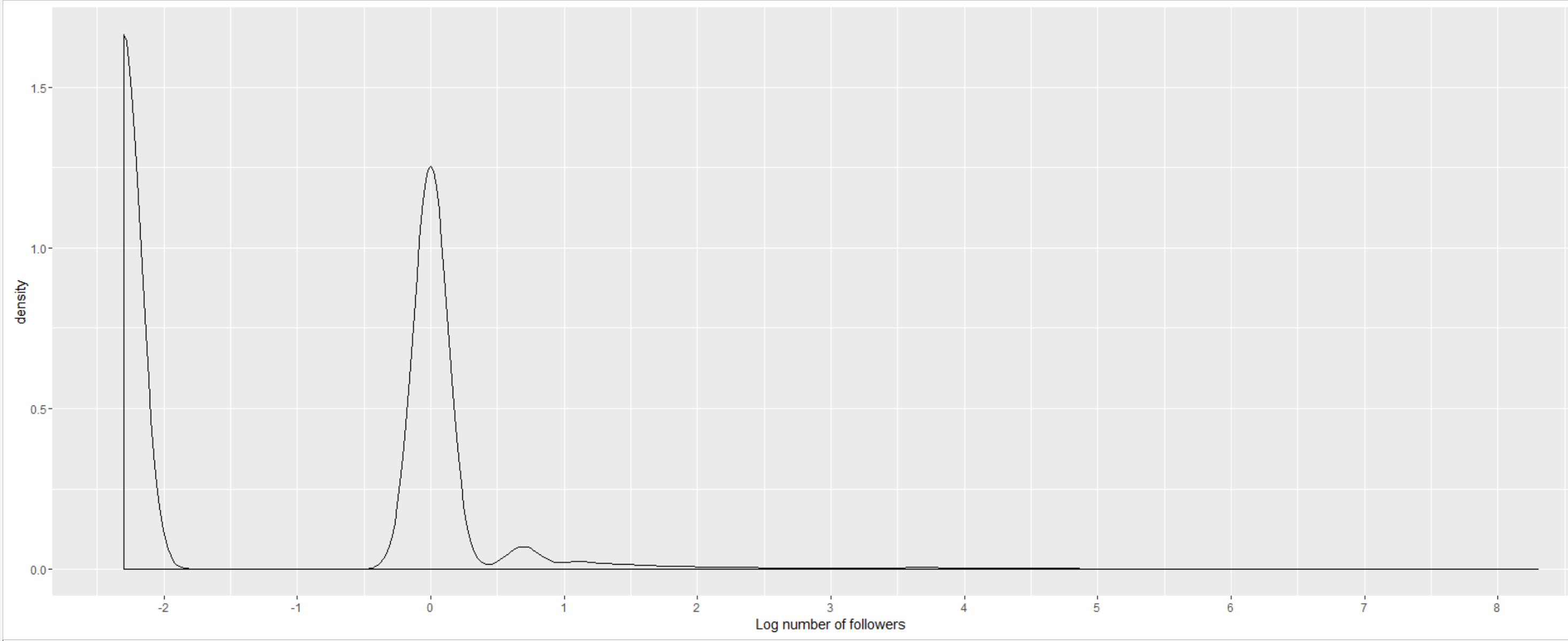
To plot the distribution of the number of followers, we use R and ggplot2 library to perform necessary calculations:



We calculated the follower count of each node in the graph including the nodes with zero followers then we plot the “Number of followers” as the X axis against the “Node counts” as the Y axis with a bar plot.

###### 2. Transform distribution of the of number of followers to log scale

We can see from the previous plot that the distribution is highly skewed. A log transformed distribution can help use to get a better understanding of the distribution. To properly perform log transformation, we replace 0 with 0.1 for the case of zero followers. A density plot is produced after log transformation:



From the log transformed distribution, we can see that there is a large amount of accounts with zero follower (exp(-2.302585) = 0.1 – log transformed zero followers) and many accounts with one follower (exp(0) = 1).

###### 3. The average number of followers and the standard deviation

R psych library provides convinent function to calculate the mean and standard deviation of the followers’ data set.

> describe(followerDist$follower\_counts)

vars n mean sd median trimmed mad min max range skew kurtosis se

1 59973 1.22 33.25 0 0.45 0 0 4020 4020 117.78 14223.51 0.14

From the data set, we can see that on average a twitter user has 1.22 follower and the standard deviation is 33.25. Although the average followers count is relative small, but there is a very large variation between different twitter accounts. From the data set we have also learned that there is a large amount of accounts that have zero followers.

###### 4. The Twitter usernames of the top 10 users with the most followers

Below are the output of top 10 users with the most followers

user\_id follower\_counts name

1 0 4020 Snapchat

2 1 4020 insomniacevents

11113 2 4020 Dropbox

22224 3 4020 olympiacos\_org

1899 11705 120 elcapimar

3203 1288 120 Kill\_Joy7

5534 14978 120 JohnIrons95

8937 1804 120 sydneyhbrodsky

14859 2337 120 Can1ffs\_bae

15226 2370 120 mcclainxkylie