A STUDY OF NETWORK PARADOXES

By

SNEHA AJIT PILGAONKAR

Bachelor of Engineering in Computer Science

Mumbai University

Mumbai, Maharashtra, India

2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December, 2014

A STUDY OF NETWORK PARADOXES

Thesis Approved:

Dr. Subhash Kak

Thesis Adviser

Dr. Christopher Crick

Dr. David Cline

Name: SNEHA AJIT PILGAONKAR

Date of Degree: DECEMBER 2014

Title of Study: A STUDY OF NETWORK PARADOXES

Major Field: COMPUTER SCIENCE

Abstract:

Social networks are characterized by their user-oriented nature, and interactive, community-driven and emotion based content. There are some known paradoxes about social network of which ‘friendship paradox’ is the most famous one. According to this paradox, on an average your friends have more friends than you have. This paradox can have either of the two types of origin: statistical or behavioral. A statistical origin can be determined based on the mathematical properties of social connectivity such as mean and median. The research analyzes this problem by using data from one of the online social networks. We first show that the social connectivity data does not satisfy Benford’s law. This fact and other statistical analysis performed by us establish that the friendship paradox data from social networks has a large behavioral component.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

1.1 Network paradoxes

The basic components of a network are the nodes and edges. A social network is constituted of interconnected nodes and edges where the nodes are the uses and the edges represent the connectivity between each pair of users. The interaction between users and their decision of whom to link with leads to the formation of the social network.

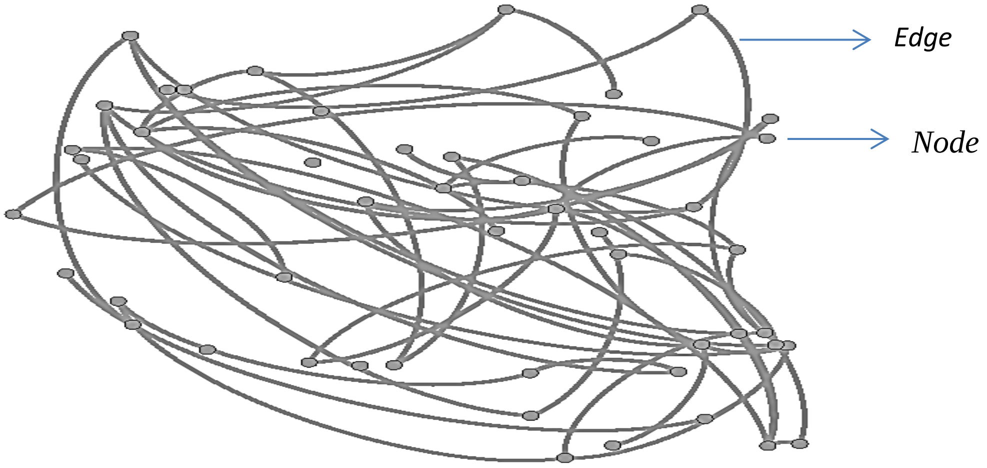


Figure 1.1 Structure of Social Network

For example, in Facebook, whom you add as friends or whom you follow forms the topology of the network.

Networks are associated with some properties that may appear paradoxical at first sight. Some of the prominent paradoxes are given below:

I. *Braess's Paradox*. This paradox is named after Dietrich Braess which states that adding extra capacity to a moving network when the moving entities rationally choose their route can reduce overall performance [1]. The Braess Paradox has relevance in traffic control management. Consider the following example :

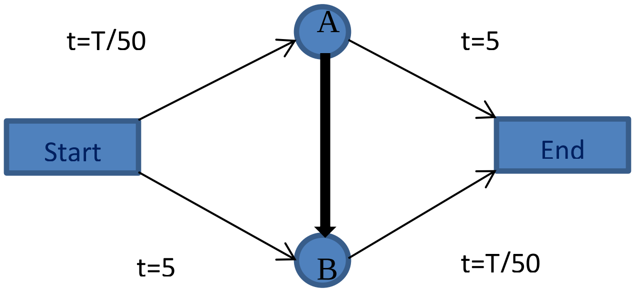


Figure 1.2 Braess paradox network [35]

In the road network of Figure 1.2, suppose if 200 drivers are travelling from *Start* to *End*. On the road *Start-A* the time taken to travel is t= (number of travelers) divided by 100 i.e. t=200/50 = 4 whereas on the *Start-B* time taken is t= 5 which is constant and similarly on the other cross roads. Consider if the road connecting *A* and *B* doesn't exist. The time taken from *Start-A-End* would be 4+5=9 and from *Start-B-End* would be 5+4=9. If the system is supposed to be in Nash Equilibrium then out of 200 drivers, 100 drivers would take the *A* route while other half would take *B* route. Hence each route takes 7 minutes to reach the *End*.

Now consider the road connecting *A* and *B* has negligible travelling distance. The drivers will now choose the shortest route *Start-A* route which will require maximum of 4 minutes whereas *Start-B* takes constant of 5 minutes. From point *A* the drivers will select the *A-B route* and from there similarly *B-End* route which adds up the travel time to 4+4=8 minutes which is greater than 7 minutes required when *A-B* did not exist. Hence it shows that if every driver selects the road which seems to be most favorable then he will end up in not choosing the road requiring optimal travelling time.

Let $T_e(x)$ be the time required for *x* drivers driving along edge *e* [2]. If the graph is linear then the let energy edge *e* for *x* drivers, E(e) be :

$\sum_{i=1}^{e^x} Te(i)$ = $T_e(1) + T_e(2) +$ ….. $+ T_e(e^x)$

The total energy of the graph is the sum of energies of every edge. Consider that the system is not in equilibrium and one of the drivers decides to opt for the most favorable path. If the original path is along the edges $a_1, a_2,$ ….. a_n and the new path is $b_1, b_2,$ ….. $b_n.$ After choosing the new path, the old one is removed and new one is added. Hence the energy associated with the old one for the entire graph *E* will be reduced by $\sum_{i=0}^{n} T_e (e^x)$. Since the new path is shorter than the old one, the energy *E* will reduce. If the process is repeated, *E* will decrease further and since *E* has to be positive equilibrium will occur.

II. *Parrondo's Paradox.* This paradox is named after Parrondo who discovered it in 1996. The paradox states that there exist pairs of games, each with a higher probability of losing than winning, for which it is possible to construct a winning strategy by planning the games alternately [3].

III. *Friendship Paradox.* The friendship paradox was discovered by sociologist Scott Feld in 1991 which states that “regardless of which individual we pick in the social network, on average, his/her friends have more friends than he/she has” [4]. This paradox is normally contingent on the existence of symmetric relation between users. In Facebook, for example if A is friend of B then obviously B is also friend of A which marks the symmetric relation. But this paradox also appears to be applicable to social networking sites like Twitter where the users follow each other [5].

A statistical fact references in PEW Internet and American Project 2011 is that the average user in Facebook has 245 friends but the average friend of that user has 359 friends [6]. The intuitive explanation of this is that there are small sets of people isolated in a network and hence they do not appear in most peoples’ network. But the popular people appear repeatedly and hence on average your friends always appear to have more friends than you. But as we will see later, the mathematical proof is more subtle than this explanation.

1.2 Origin of Network Paradoxes

Statistical sampling gives rise to paradox in mean. For example in Twitter, the user’s attribute of mean of number of followers compared to the mean of the followers of the user’s followers results in paradox even if network is completely random [4]. In any social network, the paradox exists on mean for any attribute. Such paradox is also called as ‘Traditional paradox’. It has been found that the paradox also exists in median for social network. This means that the paradox is not just the result of statistical properties of network but has a behavioral origin.

The behavioral origin of paradoxes has been proved by shuffling the network which destroys all the correlations between the nodes and correlation within the nodes and its attributes. Shuffling the network means interchanging the number of attributes of a node with another node but its distribution in the network remains same.

The main objective of our work is to establish that the friendship paradox for social network has a large behavioral component. In the examination of its mathematical origin, we use statistical parameters of *mean* and *median*.

CHAPTER II

REVIEW OF LITERATURE

2.1 Previous work for statistical and behavioral origin of network paradoxes

As discussed in the previous chapter, social network paradoxes are not purely statistical but they also arise due to behavioral reasons. In order to prove that the network paradox has behavioral origin, one study examined the effect of the removal of correlation on it [4]. This study performed shuffle test that involves interchanging the values of attributes of a nodes and assigning new one to all. The network links between the nodes remain the same and hence the overall distribution of the network remains unchanged.

For example if the node attribute is number of friends of a user, new friends are added for a node which are taken from other neighboring nodes. Due to this shuffling of network, the assortativity is destroyed but friendship paradox exists in the mean but ceases to exist in median. Shuffling results in both the correlations existing in the network to be destroyed. The origin of strong network paradox exists in the between and within-node correlations.

After shuffling, paradox exists in mean but it cannot be determined if it is due to within-node or between-node correlation. Therefore to sort this confusion controlled shuffle test was performed in which at a given time within-node correlation is eliminated and between-node correlation still exists and vice-a–versa was also performed.

Controlled shuffle test is achieved by grouping that node together which has same value for degree attribute. After grouping the nodes, shuffling is carried out in each group. Sometimes both type of correlation exists because of the nature of the attributes of the nodes in the network. Considering within-node correlation we can see that if one of the attribute likes ‘number of friends’ change or exists because users add more friends or delete friends and eventually affects the existence of the correlation. On the other hand for between-node correlation assortativity takes place i.e. nodes place themselves near those nodes that have matching attributes.

A social network is characterized by heavy-tailed degree distribution. Due to the largeness of social network data, sampling techniques for heavy-tailed degree distributions using fraction of sample nodes are needed. The authors in [7] proposed a new method called ‘tail-scope’ which has more advantages over the traditional uniform node sampling method for estimating heavy tailed degree distributions [8]. Later a hybrid method comprising of both the UNS and ‘tail-scope’ was devised. This method helps in getting the structure of network by estimating the heavy-tailed degree distribution.

Friendship paradox proves to be a useful tool to estimate heavy tails in distribution. The higher degree nodes are much more likely to be observed by their respective neighbors. Using this observation, sampling of nodes is carried out. The results show that from randomly chosen nodes with group of friends have more highly connected nodes than those formed by uniformly sampled nodes [7]. These ideas have also been applied to a growing network, mechanism followed by *degree* and *qualities.*

2.2 Statistical and Behavioral properties of social network:

We will see later that the friendship paradox has its foundation in the heterogeneous distribution of nodes attributes like node degree. The paradox with mean values of user's traits arises from statistical sampling from a heavy-tailed distribution. Attributes such as node degree often have a heavy tail when it has extremely large values, which means that the user who is very popular appears much more frequently than expected.

Degree is the most common attribute for a node in social network for which we get heavy-tailed distributions. Degree in a network is the number of other nodes that a particular node is connected. Apart from degree many other attributes exist for which we do get heavy-tailed distribution such as activity, diversity, virality etc.

For social network sites, a stronger version of friendship paradox holds known as *Strong friendship paradox* [9]. Strong friendship paradox states that majority of your friends have more friends than you do.

Behavioral nature is related with the correlation between nodes and correlation within node and attributes. There are two types of correlation in a network: between-node correlation and within-node correlation:

a) Within-node correlation- This type of correlation exists between users' degree and its own attributes. Pearson's Correlation Co-efficient is used to measure this correlation.

b) Between-node correlation- This type of correlation exists between attributes of a node and attributes of its neighbors. Assortativity is used to measure this correlation.

CHAPTER III

MATHEMATICAL BASIS OF FRIENDSHIP PARADOX

3.1 Degree Distribution in social network:

In a social network we know that some nodes have a high degree while some have less value of the degree. Since it is driven by the community, variations in degree of nodes are to be expected [10]. When sampling is carried out, there will always be uneven tails on both sides or we can say that the distribution is skewed or not symmetric. For a skewed distribution, the values of mean and median are not same. Skewness pulls the mean is pulled in that direction. When the right tail is heavier then mean is greater than median and for heavy left tail mean is less than median. In Figure 3.1 we can see that the right tail is heavier than the left tail.

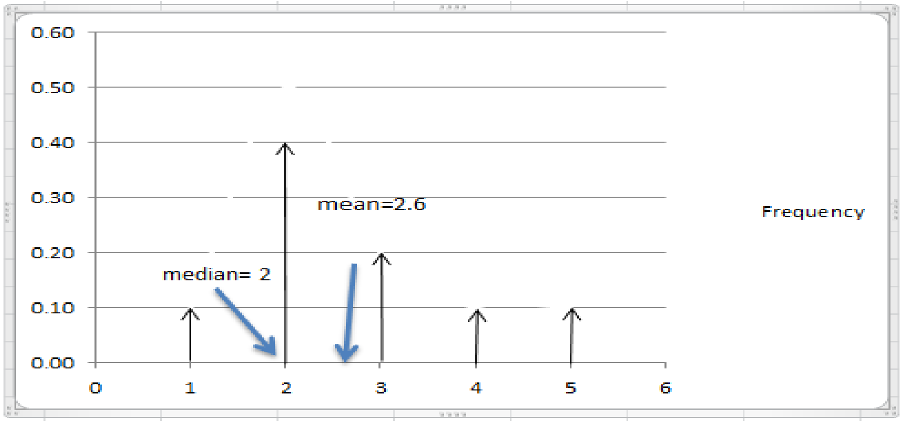


Figure 3.1 Distribution with right heavy tail

Due to which the mean is shifted towards the heavy tail i.e. the right side resulting in the value of mean (2.6) being greater than value of median (2.0). In the above graph, we see that there are very few nodes with higher probability and more nodes with lower probability. The nodes with higher probability skew the average and pull the value towards the side of heavy tail. In the above graph, it has right heavy tail and hence mean is greater that median. In Figure 3.2 we can see that the left tail is heavier than the right tail where the mean (4.375) is less than median (5).

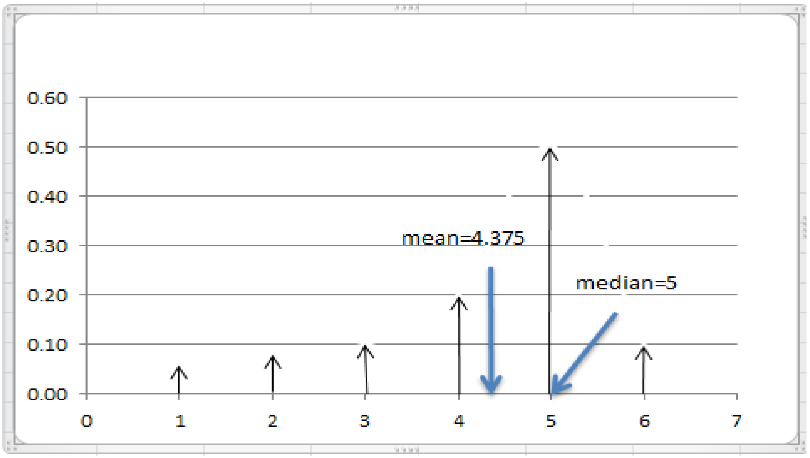


Figure 3.2 Distribution with left heavy tail

3.2 Examples of Friendship Paradox:

In order to understand how exactly the paradox is true, we will take a small example of four individuals (Figure 3.3).

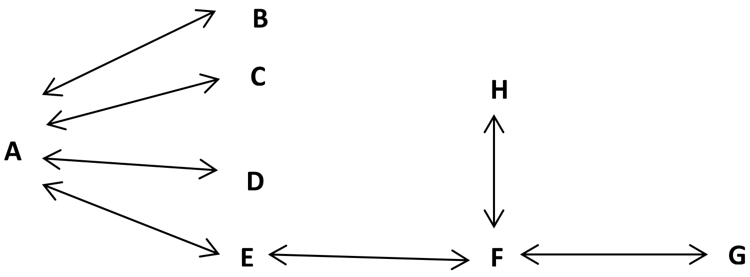


Figure 3.3 Example network-1

Let there be four individuals A, B, C and D. We need to basically compare the average number of friends of an individual to the average number of friends of the friends of that individual.

In the above graph also, we see that there are very few nodes with higher probability and more nodes with lower probability. The nodes with higher probability skew the average and pull the value towards the side of heavy tail. In the above graph, it has left heavy tail and hence mean is less that median.

In the Figure 3.3 graph, the following is true:

A has 4 friends – B, C, D, E

B has 1 friend – A

C has 1 friend - A

D has 1 friend - A.

E has 2 friends – A and F

F has 3 friends – E, G, H

G has 1 friend - F

H has 1 friend - F

Now we know that A has 4 friends B, C, D and E of which B has 1, C has 1, D has 1 and E has 2 friends. Hence total friends of A become 5. Similarly total friends for B become 4, for C friends are 4, for D friends are 3, for E friends are 7, for F friends are 4, for G friends are 3 and for H friends are 3

The next step is to calculate the average number of friends of each individual's friends which is calculated by dividing the two values that we calculated above for each individual.

Calculating average number of friends of each individual's friends

A- 1.25
B- 4.00
C- 4.00
D- 4.00
E -3.50
F – 1.33
G- 3.00
H- 3.00

Taking a look at the above values we see that the average number of friends of each individual's friends for A is the least which an exception because A is the most famous individual with highest node degree i.e. higher number of friends. Summarizing the above calculation, we can present the date in the form of the table below.

	#friends	#friends_of_friends	Average_friends_of_friends
A	4	5	1.25
B	1	4	4.00
C	1	4	4.00
D	1	4	4.00
E	2	7	3.50
F	3	4	1.33
G	1	3	3.00
H	1	3	3.00

Figure 3.4 Table showing the calculation for Friendship paradox for network-1

From the above table, we get μ for individual user which is 1.75 whereas μ of friends of friend of the user is 2.43. This means that average individual has 1.75 friends but the average friend has 2.43 friends. Hence we get that, on average your friends have more friends than you have.

Let us see one more example for friendship paradox.

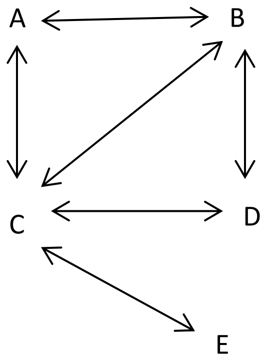


Figure 3.5 Example network-2

Consider the above network graph in which we can see that among the five individuals.

A has 2 friends – B, C

B has 3 friends – A, C, D

C has 4 friends – A, B, D and E

D has 2 friends – B, C

E has 1 friend - C

Now we know that A has 2 friends B and C of which B has 3 and C has 4. Hence total friends of A become 7. Similarly total friends for B become 8, for C friends are 8, for D friends are 7, for E friends are 4.

The next step is to calculate the average number of friends of each individual's friends which is calculated by dividing the two values that we calculated above for each individual.

Calculating average number of friends of each individual's friends

A- 3.5
B- 2.66
C- 2
D -3.5
E – 4

Taking a look at the above values we see that the average number of friends of each individual's friends for A is the least which an exception because A is the most famous individual with highest node degree i.e. higher number of friends.

Summarizing the above calculation,

	#friends	#friends_of_friends	Average_friends_of_friends
A	2	7	3.5
B	3	8	2.66
C	4	8	2.00
D	2	7	3.50
E	1	4	4.00

Figure 3.6 Table showing the calculation for Friendship paradox for network-2

From the above table, we get μ for individual user which is 2.4 whereas μ of friends of friend of the user is 2.83. This means that average individual has 2.4 friends but the average friend has 2.83 friends. Hence we get that, on average your friends have more friends than you have.

3.3 Mathematical Proof of Friendship Paradox

Considering the friendship paradox, we will need to find the average number of friends for an entire graph and the average number of friends of an individual's friends [11].

We will follow mathematical approach in order to prove the friendship paradox. As said above we need two values – the average friends and the average of friends' friends.

Let *i* be the representation for any individual for e.g. A, B, C or D and let *t* denote the total number of individuals. Hence for any individual i, total number of friends are $f_{i.}$

Now we can calculate the average number of friends in the network which is as follows,

$$\mu_f = \sum \frac{fi}{n}$$

where μ_f is the average number of friends in network, f_i is the number of friends and n is the total number of individuals in the network.

Now we calculate the second quantity i.e. average number of friends of friends in the network. For this we need to calculate friends for each individual and finally sum them up. For each individual *i* (A, B, C or D) his/her total friends are the sum total of friends of his/her friends.

For example, if *i* is B then friends of B according to Figure 2 are A and D. Hence it would be the sum of f_A + f_D. Similarly for A, it would be f_B + f_C + f_D.

We notice that for any individual *i* we need to include his count of number of friends f_i in the final summation when his friends are considered.

In the above calculation we see that for any term f_i it is included in the sum only when his/her friends f_i are considered.

This gives us that for every individual *i* ; the final summation includes the term

$$(fi)\ \ (fi) = fi^2$$

We get the total number of friends of an individual's friends by summation of = $\sum f_{\mathrm{i}}{}^2$

The average number of friends of an individual's friends is calculated by the ratio of total

$$= \frac{\text{total number of friends of an individual's friends}}{\text{total number of individual's friends.}}$$

So we get average number of friends of an individual's friends as $\mu_{ff} = \sum f_{\mathrm{i}}{}^2 \,/\, \sum f_{\mathrm{i}}$

The formula for variance (σ) is given by- $\sigma^2 = (\sum f_{\mathrm{i}}{}^2 / \mathrm{n}) - \mu^2$

Simplifying it further, we get $\sum f_{i}{}^{2} = (\mu^2 + \sigma^2)\ n$

Dividing L.H.S by $\sum f_{i}$ and R.H.S by μ_n (since $\sum f_{i} = \mu_n$), we get

$$\boldsymbol{\mu}_{ff} = \sum f_{i}{}^{2} \;/\; \sum f_{i} \quad = (\mu^2 + \sigma^2)\ n \;/\; \mu n \quad = \mu + (\sigma^2 / \mu) \;\ldots\ldots\ldots\ldots\; \text{(I)}$$

Finally we get the mathematical formula for average number of friends of an individual's friends.

Let us compare the two values average number of friends μ and average number of friends of an individual's friends $\mu + (\sigma^2 / \mu)$.

As we can observe, $\boldsymbol{\mu_f \leq \mu_{ff}}$ which proves that the average/mean number of friends of an individual is less than or equal to the average/mean number of friends of his/her friends. This mathematical proof proves the Friendship paradox.

In the above proof in line (I), we get $\mu_{ff} = \mu + (\sigma^2 / \mu)$

As stated in the PEW Internet and American Project 2011, average friends of friends for any user are 359 and average friends of the user are 245 [6].

$\mu + (\sigma^2 / \mu) = 359$

$245 + (\sigma^2 / 245) = 359$

Solving further we get, $\sigma^2 = \sqrt{}\ 27930 =$

Hence, $\sigma = 167.112$

We can say that since σ has a large value, the data points are spread out around the mean in the heavy tail distribution.

The Friendship Paradox is not just related with degree of count of friends but it can be stated in the following forms:

a) On average your friends have more friends than you do.

b) On average your followers have more friends than you do.

c) On average your friends have more followers than you do.

d) On average your followers have more followers than you do.

Other similar paradoxes also exist such as Twitter has users who tweets fewer messages and receive less-viral content than his/her friends do on average, A scientist’s co-authors are more productive and better cited on average [5]. Network paradoxes result in systematic biases in how individuals perceive their world.

CHAPTER IV

STATISTICAL ANALYSIS OF DATA

4.1 Data Analysis using R

R is an open source language implemented from S (Statistical Language). It has more than 4000 packages catering to various applications like exploratory analysis, graphics and visualization, cluster analysis, natural language processing etc. [12].

R offers a special package for Facebook data analysis called ‘Rfacebook’. This package offers variety of functions which can be utilized to access Facebook’s API to retrieve information about Facebook user’s details and other updates. In order to access any API requests, an access token is required for the same. There are two ways to acquire an access token, one which produces a temporary access token whereas the other produces access token for a much longer duration

```
# token generated here: https://developers.facebook.com/tools/explore
```

```
Token <─ “access-token”
```

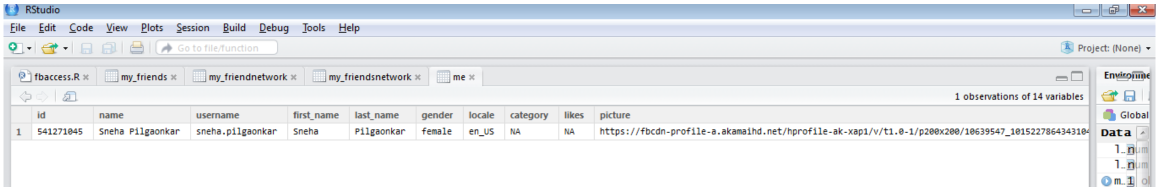
Access token granted by Facebook is used by methods in R for further process. A function ‘fbOauth’ is used which takes ‘appid’ and ‘appsecret’ as input and authorizes the user for Facebook.

Fb_outh<─fbOAuth(app_id="596561593788794",app_secret="cac820fb2c8acdd83d6e5540410c9442",extended_permission = T)

Further we get the information of the user, friends list of the user and his friends network using the functions ‘getUsers’,’getFriends’,’getNetwork’ respectively.

4.2 Retrieving Facebook data

me <- getUsers("snehapilgaonkar”, token,private_info=TRUE)

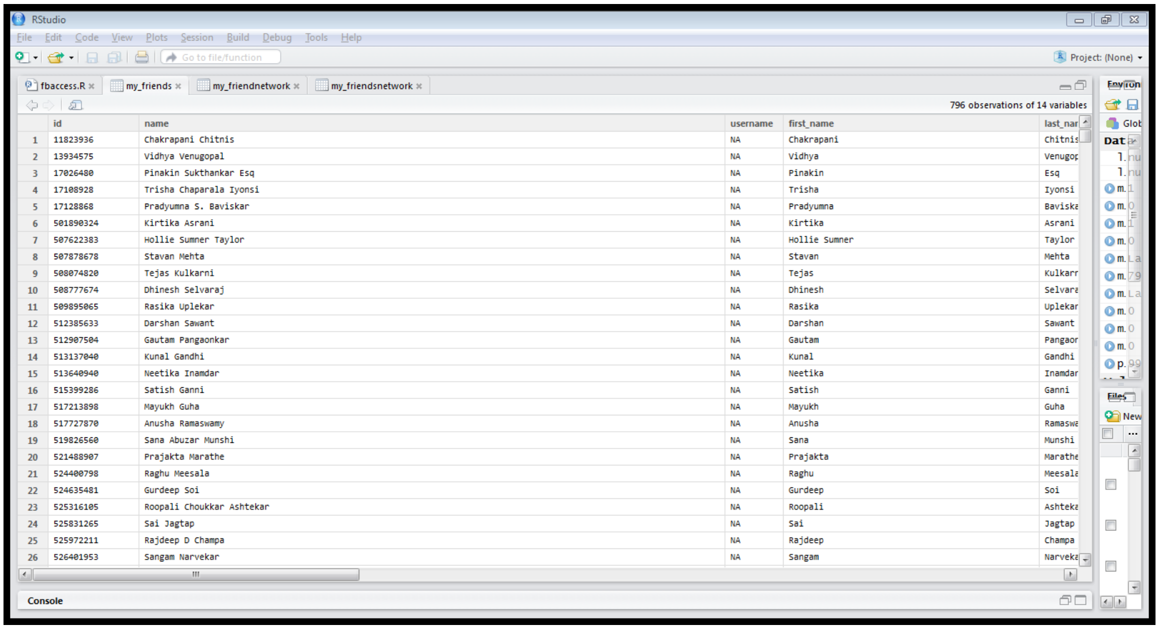


1 observations of 14 variables

	id	name	username	first_name	last_name	gender	locale	category	likes	picture
1	541271045	Sneha Pilgaonkar	sneha.pilgaonkar	Sneha	Pilgaonkar	female	en_US	NA	NA	https://fbcdn-profile-a.akamaihd.net/hprofile-ak-xap1/v/t1.0-1/p200x200/10639547_1015227864343104

Figure 4.1 R command -getUsers

my_friends <- getFriends(token, simplify = F)

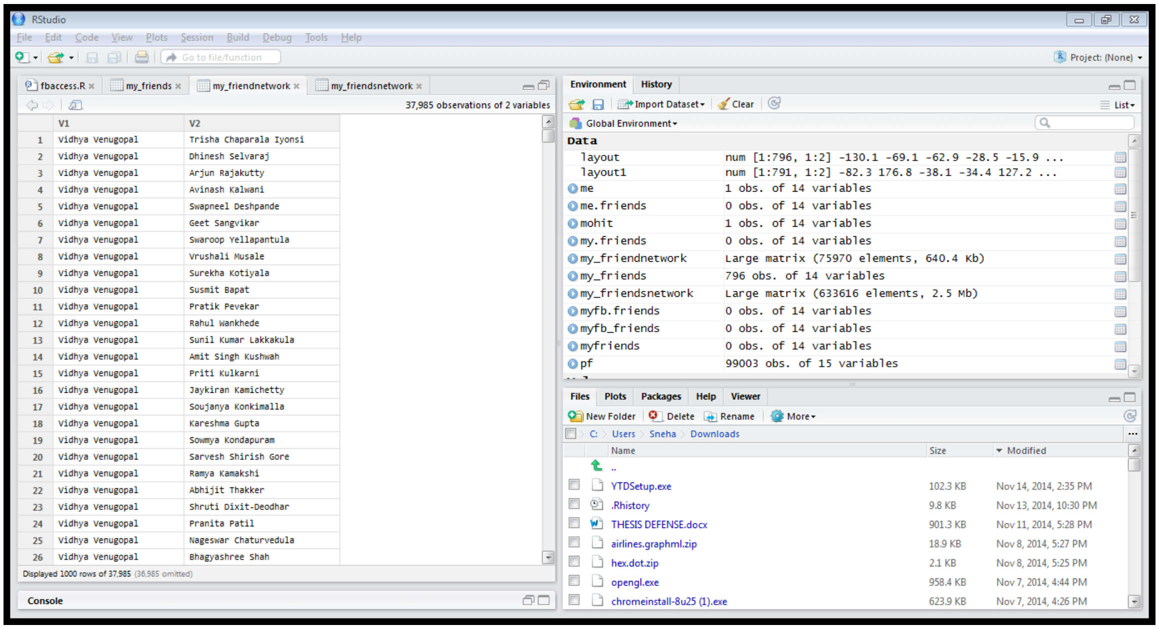


796 observations of 14 variables

	id	name	username	first_name	last_nar
1	11823936	Chakrapani Chitnis	NA	Chakrapani	Chitnis
2	13934575	Vidhya Venugopal	NA	Vidhya	Venugop
3	17026480	Pinakin Sukthankar Esq	NA	Pinakin	Esq
4	17108928	Trisha Chaparala Iyonsi	NA	Trisha	Iyonsi
5	17128868	Pradyumna S. Baviskar	NA	Pradyumna	Baviska
6	501890324	Kirtika Asrani	NA	Kirtika	Asrani
7	507622383	Hollie Sumner Taylor	NA	Hollie Sumner	Taylor
8	507878678	Stavan Mehta	NA	Stavan	Mehta
9	508074820	Tejas Kulkarni	NA	Tejas	Kulkarr
10	508777674	Dhinesh Selvaraj	NA	Dhinesh	Selvara
11	509895065	Rasika Uplekar	NA	Rasika	Uplekar
12	512385633	Darshan Sawant	NA	Darshan	Sawant
13	512907504	Gautam Pangaonkar	NA	Gautam	Pangaor
14	513137040	Kunal Gandhi	NA	Kunal	Gandhi
15	513640940	Neetika Inamdar	NA	Neetika	Inamdar
16	515399286	Satish Ganni	NA	Satish	Ganni
17	517213898	Mayukh Guha	NA	Mayukh	Guha
18	517727870	Anusha Ramaswamy	NA	Anusha	Ramaswa
19	519826560	Sana Abuzar Munshi	NA	Sana	Munshi
20	521488907	Prajakta Marathe	NA	Prajakta	Marathe
21	524400798	Raghu Meesala	NA	Raghu	Meesala
22	524635481	Gurdeep Soi	NA	Gurdeep	Soi
23	525316105	Roopali Choukkar Ashtekar	NA	Roopali	Ashteka
24	525831265	Sai Jagtap	NA	Sai	Jagtap
25	525972211	Rajdeep D Champa	NA	Rajdeep	Champa
26	526401953	Sangam Narvekar	NA	Sangam	Narveka

Figure 4.2 R command- getFriends

my_friendnetwork <— getNetwork(token, format="edgelist", verbose=TRUE)



37,985 observations of 2 variables

	V1	V2
1	Vidhya Venugopal	Trisha Chaparala Iyonsi
2	Vidhya Venugopal	Dhinesh Selvaraj
3	Vidhya Venugopal	Arjun Rajakutty
4	Vidhya Venugopal	Avinash Kalwani
5	Vidhya Venugopal	Swapneel Deshpande
6	Vidhya Venugopal	Geet Sangvikar
7	Vidhya Venugopal	Swaroop Yellapantula
8	Vidhya Venugopal	Vrushali Musale
9	Vidhya Venugopal	Surekha Kotiyala
10	Vidhya Venugopal	Susmit Bapat
11	Vidhya Venugopal	Pratik Pevekar
12	Vidhya Venugopal	Rahul Wankhede
13	Vidhya Venugopal	Sunil Kumar Lakkakula
14	Vidhya Venugopal	Amit Singh Kushwah
15	Vidhya Venugopal	Priti Kulkarni
16	Vidhya Venugopal	Jaykiran Kamichetty
17	Vidhya Venugopal	Soujanya Konkimalla
18	Vidhya Venugopal	Kareshma Gupta
19	Vidhya Venugopal	Sowmya Kondapuram
20	Vidhya Venugopal	Sarvesh Shirish Gore
21	Vidhya Venugopal	Ramya Kamakshi
22	Vidhya Venugopal	Abhijit Thakker
23	Vidhya Venugopal	Shruti Dixit-Deodhar
24	Vidhya Venugopal	Pranita Patil
25	Vidhya Venugopal	Nageswar Chaturvedula
26	Vidhya Venugopal	Bhagyashree Shah

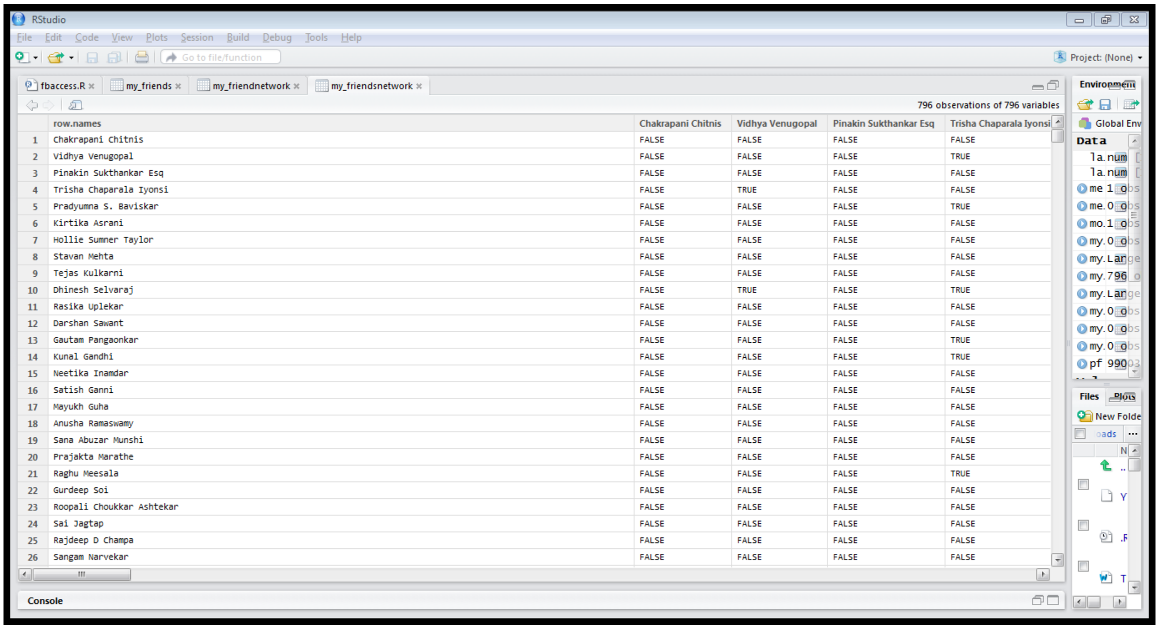
Displayed 1000 rows of 37,985 (36,985 omitted)

```
Data
  layout              num [1:796, 1:2] -130.1 -69.1 -62.9 -28.5 -15.9 ...
  layout1             num [1:791, 1:2] -82.3 176.8 -38.1 -34.4 127.2 ...
  me                  1 obs. of 14 variables
  me.friends          0 obs. of 14 variables
  mohit               1 obs. of 14 variables
  my.friends          0 obs. of 14 variables
  my_friendnetwork    Large matrix (75970 elements, 640.4 Kb)
  my_friends          796 obs. of 14 variables
  my_friendsnetwork   Large matrix (633616 elements, 2.5 Mb)
  myfb.friends        0 obs. of 14 variables
  myfb_friends        0 obs. of 14 variables
  myfriends           0 obs. of 14 variables
  pf                  99003 obs. of 15 variables
```

Name	Size	Modified
..		
YTDSetup.exe	102.3 KB	Nov 14, 2014, 2:35 PM
.Rhistory	9.8 KB	Nov 13, 2014, 10:30 PM
THESIS DEFENSE.docx	901.3 KB	Nov 11, 2014, 5:28 PM
airlines.graphml.zip	18.9 KB	Nov 8, 2014, 5:27 PM
hex.dot.zip	2.1 KB	Nov 8, 2014, 5:25 PM
opengl.exe	958.4 KB	Nov 7, 2014, 4:44 PM
chromeinstall-8u25 (1).exe	623.9 KB	Nov 7, 2014, 4:26 PM

Figure 4.3 R command – getNetwork(edgelist)

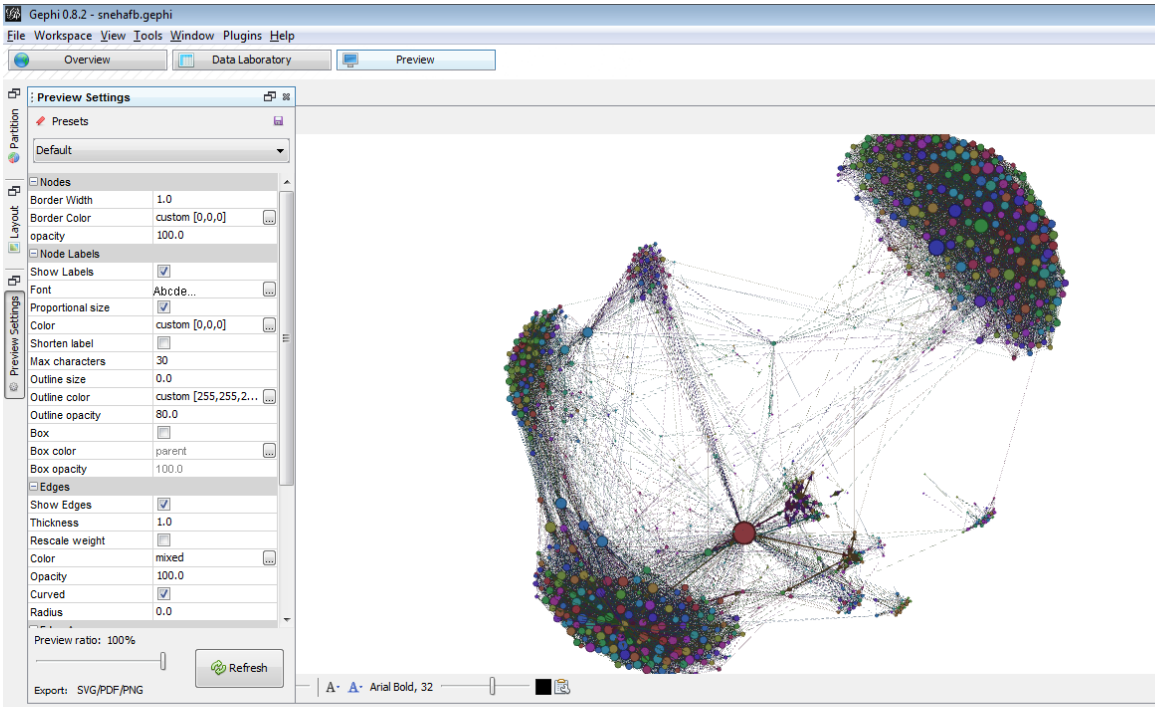
my_friendsnetwork<—getNetwork(token,format="adj.matrix",verbose=TRUE)



796 observations of 796 variables

	row.names	Chakrapani Chitnis	Vidhya Venugopal	Pinakin Sukthankar Esq	Trisha Chaparala Iyonsi
1	Chakrapani Chitnis	FALSE	FALSE	FALSE	FALSE
2	Vidhya Venugopal	FALSE	FALSE	FALSE	TRUE
3	Pinakin Sukthankar Esq	FALSE	FALSE	FALSE	FALSE
4	Trisha Chaparala Iyonsi	FALSE	TRUE	FALSE	FALSE
5	Pradyumna S. Baviskar	FALSE	FALSE	FALSE	TRUE
6	Kirtika Asrani	FALSE	FALSE	FALSE	FALSE
7	Hollie Sumner Taylor	FALSE	FALSE	FALSE	FALSE
8	Stavan Mehta	FALSE	FALSE	FALSE	FALSE
9	Tejas Kulkarni	FALSE	FALSE	FALSE	FALSE
10	Dhinesh Selvaraj	FALSE	TRUE	FALSE	TRUE
11	Rasika Uplekar	FALSE	FALSE	FALSE	FALSE
12	Darshan Sawant	FALSE	FALSE	FALSE	FALSE
13	Gautam Pangaonkar	FALSE	FALSE	FALSE	TRUE
14	Kunal Gandhi	FALSE	FALSE	FALSE	TRUE
15	Neetika Inamdar	FALSE	FALSE	FALSE	FALSE
16	Satish Ganni	FALSE	FALSE	FALSE	FALSE
17	Mayukh Guha	FALSE	FALSE	FALSE	FALSE
18	Anusha Ramaswamy	FALSE	FALSE	FALSE	FALSE
19	Sana Abuzar Munshi	FALSE	FALSE	FALSE	FALSE
20	Prajakta Marathe	FALSE	FALSE	FALSE	FALSE
21	Raghu Meesala	FALSE	FALSE	FALSE	TRUE
22	Gurdeep Soi	FALSE	FALSE	FALSE	FALSE
23	Roopali Choukkar Ashtekar	FALSE	FALSE	FALSE	FALSE
24	Sai Jagtap	FALSE	FALSE	FALSE	FALSE
25	Rajdeep D Champa	FALSE	FALSE	FALSE	FALSE
26	Sangam Narvekar	FALSE	FALSE	FALSE	FALSE

Figure 4.4 R command – getNetwork(adjacency matrix)



Preview Settings

Presets: Default

Nodes	
Border Width	1.0
Border Color	custom [0,0,0]
opacity	100.0
Node Labels	
Show Labels	☑
Font	Abcde...
Proportional size	☑
Color	custom [0,0,0]
Shorten label	□
Max characters	30
Outline size	0.0
Outline color	custom [255,255,2...
Outline opacity	80.0
Box	□
Box color	parent
Box opacity	100.0
Edges	
Show Edges	☑
Thickness	1.0
Rescale weight	□
Color	mixed
Opacity	100.0
Curved	☑
Radius	0.0

Preview ratio: 100%

Export: SVG/PDF/PNG

Figure 4.5 Gephi visualization of Facebook network

The above network graph takes nodes and edges data to create the Facebook friends' network. Gephi is used to create the visualization of the network. The clusters that can be seen represent the communities in my Facebook network. The size of the nodes depend on the degree of that node which represents my friends i.e. number of people it is connected to in the network graph.

There are some nodes which are very popular and are connected to other nodes in more than one community. The average degree and modularity is calculated and accordingly the nodes are colored. The visualization graph undergoes many layout changes such as Force atlas, Expansion, Contraction etc. which segregates the nodes, rotates the graph clockwise or anticlockwise, repels the weakly connected nodes and bring strongly connected nodes resulting in formation of clusters.

4.3 Friendship paradox on Facebook data:

Figure 4.6 consists of scatter graph of Friends-of-friends count, Mean-FF, Median-FF of the friends in a Facebook network of a circle of friends.

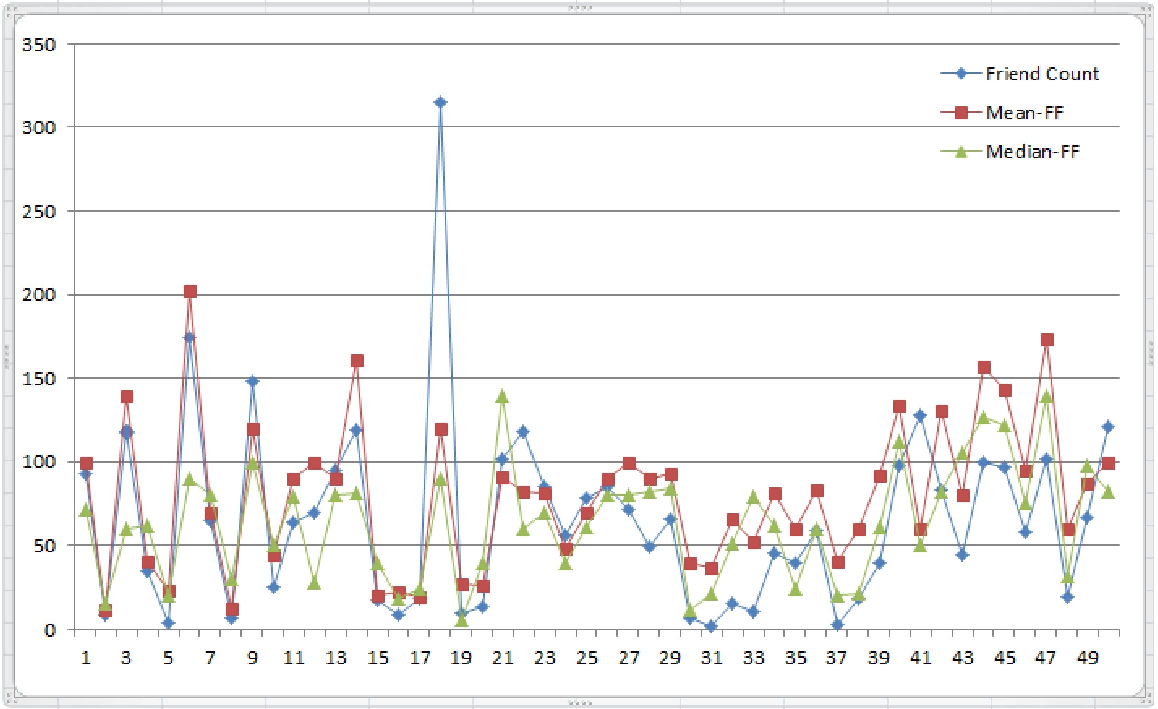


Figure 4.6 Scatter graph of Friends-of-friends count, Mean-$_{FF}$, Median-$_{FF}$ of the friends in Facebook network

As we can see in the above graph, the mean/median value is greater than that of individual’s value. Comparing the mean and median value we can observe that the friendship paradox lies in mean as well as median.

By definition friendship paradox exist in heterogeneous distribution where mean is greater than median. We can see from the graph that there lies a node with highest degree which defiantly skews the average resulting in large difference between mean and median.

The graph below clearly shows us that the mean is greater than median in almost all nodes but there are some nodes where the median is greater than the mean value. Hence friendship paradox for social network data like Facebook is not simply based on statistical values but it does have a behavioral base.

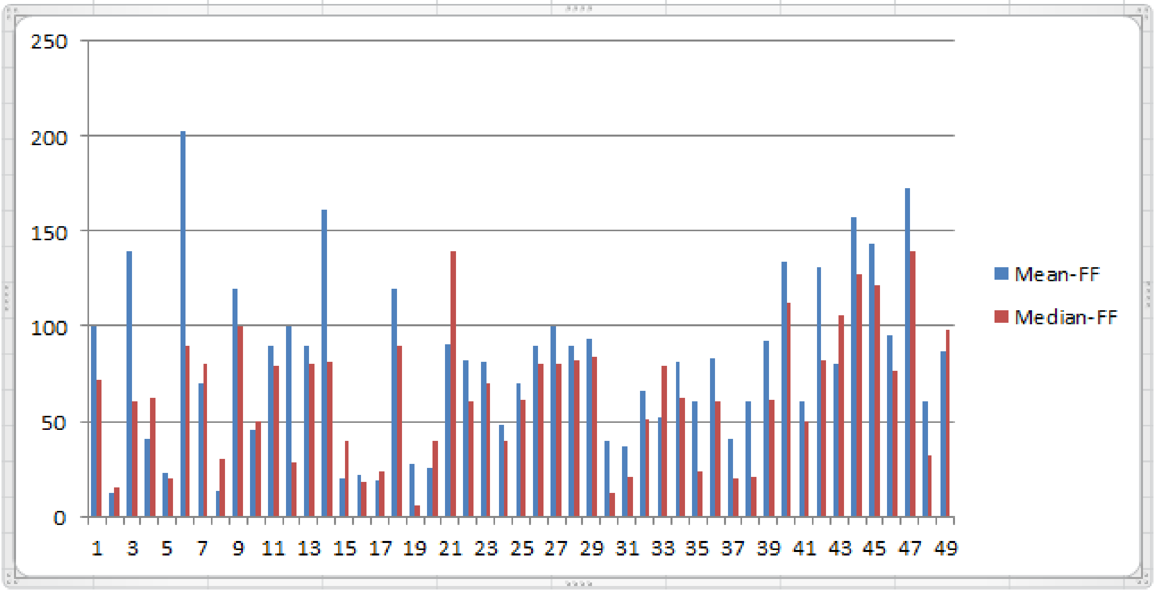


Figure 4.7 Plot of Mean-$_{FF}$ and Median-$_{FF}$ of Facebook friends

Next we compare this data to that obtained from non-social situations.

4.4 Friendship paradox on random Transportation data:

We consider a dataset of a Road Network which doesn't have any attributes which are based on human behavior or psychology. We plot the nodes and edges data in Gephi and we can see the connection of the nodes with each other. Further we analyze this data and calculate the mean and median for each node with respect to the nodes that it is connected.

We plot a graph for those values and it is seen that the mean values are always greater than the median value. We find that the paradox exists in only the mean values for the Road network.

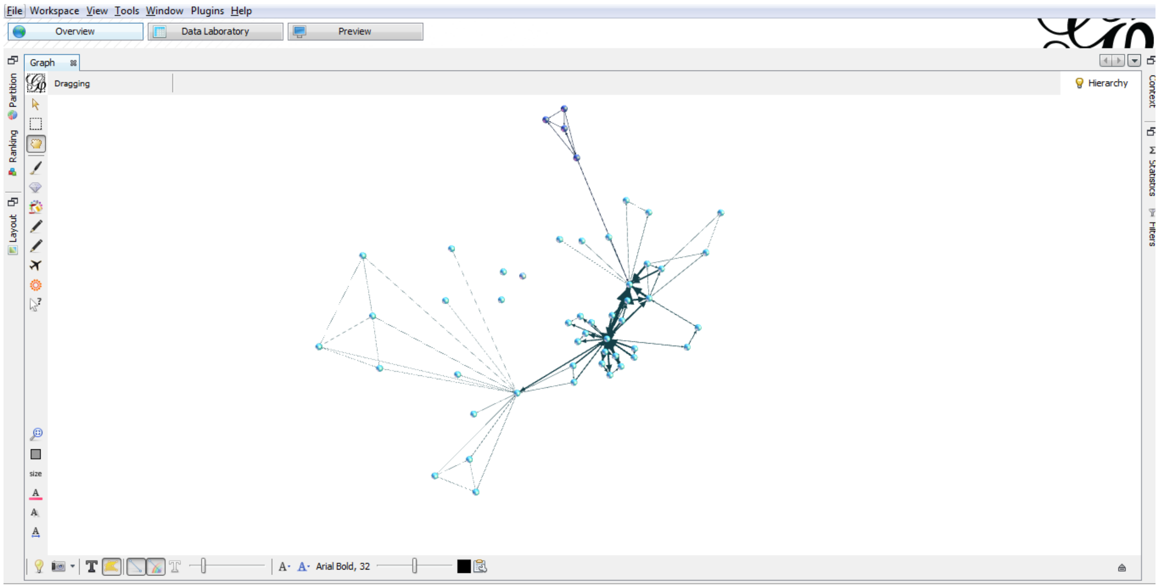


Figure 4.8 Gephi Visualization of Road Network

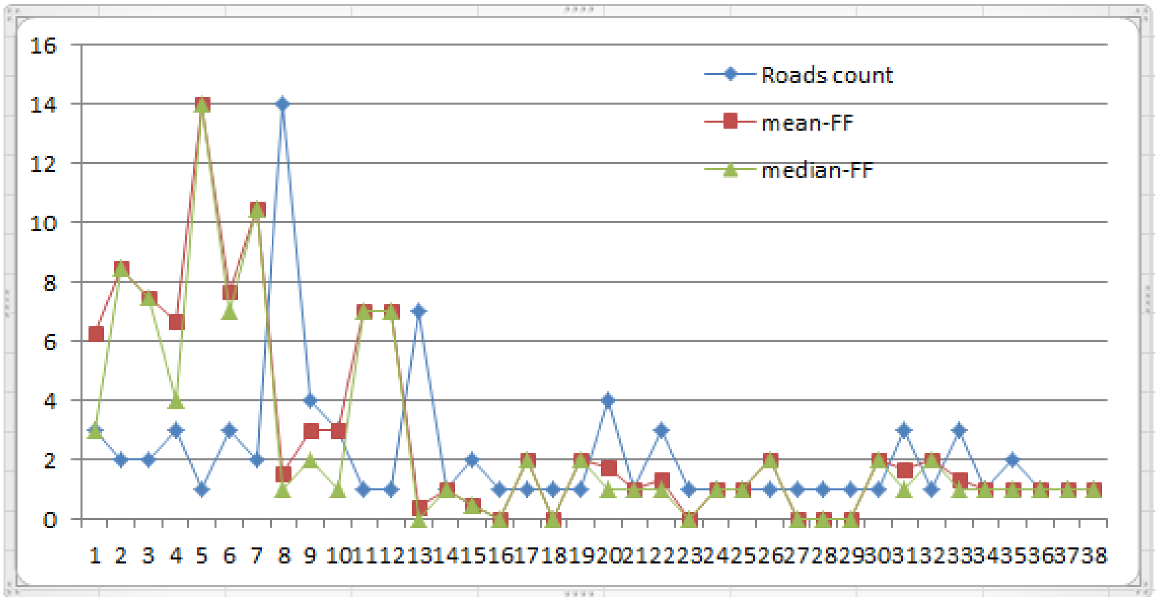


Figure 4.9 Scatter graph of degree count, Mean-$_{FF}$, Median-$_{FF}$ plot of road network

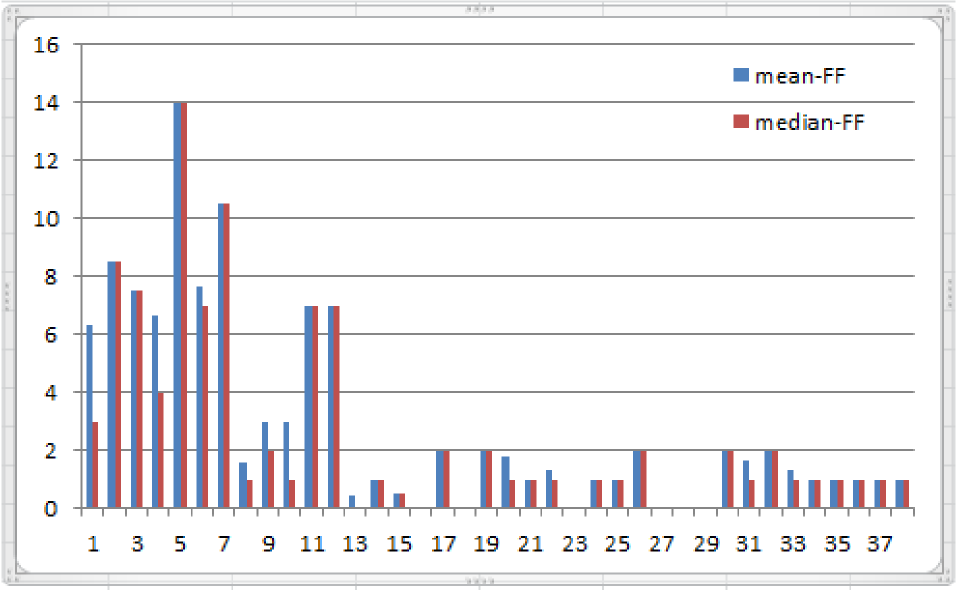
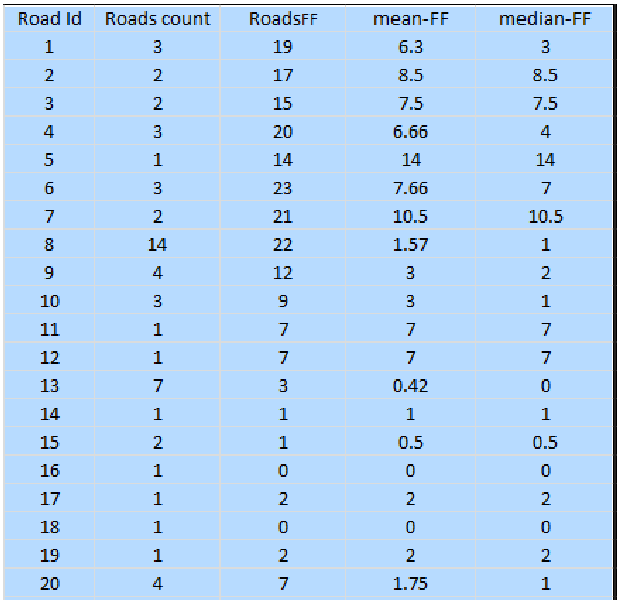


Figure 4.10 Mean-$_{FF}$ and Median-$_{FF}$ plot of nodes of Road network

Let us take a look at the actual data for the above network which will help to understand the analysis of mean$_{FF}$ and median$_{FF}$.



Road Id	Roads count	RoadsFF	mean-FF	median-FF
1	3	19	6.3	3
2	2	17	8.5	8.5
3	2	15	7.5	7.5
4	3	20	6.66	4
5	1	14	14	14
6	3	23	7.66	7
7	2	21	10.5	10.5
8	14	22	1.57	1
9	4	12	3	2
10	3	9	3	1
11	1	7	7	7
12	1	7	7	7
13	7	3	0.42	0
14	1	1	1	1
15	2	1	0.5	0.5
16	1	0	0	0
17	1	2	2	2
18	1	0	0	0
19	1	2	2	2
20	4	7	1.75	1

Figure 4.11 Mean-$_{FF}$ and Median-$_{FF}$ table of nodes of Road network

In the above table we calculated the Roads$_{FF}$, mean$_{FF}$, and median$_{FF}$ for every individual node. We observe that for the sample of nodes that we can see in the table, either mean_{FF} is greater than median_{FF} or they are equal. Hence the friendship paradox exists in mean for non-social network. On the other hand, for graph 4.10, we see that for most of the nodes the mean (FF) is greater than median (FF) values, for some nodes mean (FF) and median (FF) are equal but there are no cases in which median (FF) is greater than mean (FF). Thus for non-social network, the friendship paradox exists only in mean.

We conclude that in 29.72% of the cases in non-social network, mean (FF) is greater than median (FF). But in the sampling distribution of the Facebook (and other social network [4]), for 24% of the cases median (FF) is greater than mean (FF). This discovery is important in defining the behavioral intervention in social network paradox.

By comparing graphs for mean (FF) and median (FF) for both the data, Facebook and Road network, we can observe that the paradox for social network data has those attributes which are not purely statistical in nature.

CHAPTER V

BENFORD’S LAW- BEHAVIORAL ORIGIN

5.1 First Digit Phenomenon

In the previous chapter we saw how mean and median characteristics affect the network paradox. We found that for Facebook data, paradox exists in mean as well as median whereas for the road data it exists in mean. We are going to take a look at Benford’s Law also known as First Digit Phenomenon which is also viewed as a paradox, although it has a straightforward mathematical explanation [13].

According to Benford’s Law, tables of statistics or listings etc. the digit 1 tends to occur with the probability of approximate 30% which is much greater than the expected value 11.1% (1 digit out of 9). There are several other studies that describe the first digit phenomenon. Astronomer Simon Newcomb first observed this phenomenon in logarithmic books. He observed that the log books have their initial pages dirty i.e. seemed to have been used more often than the other page. The fellow scientists used the log tables with digit 1 more often than with digit 2 more than with digit 3 and so on.

Newcomb proposed that probability of the first significant non-zero digit can be calculated as follows:

$$\boxed{\text{Probability (first significant digit =d)} = \log_{10}\left[1+\frac{1}{d}\right] \qquad \text{(where d = 1, 2, …, 9)}}$$

Simon Newcomb had stated that the law of probability of occurrence of numbers is such that all mantissas of their logarithms are equally likely [16]. Hence for the digit 1, for mantissa 10, the probability is $\log_{10} 2$ is approximately 0.301.

In 1938, physicist Frank Benford also proposed the same logarithmic law unaware of Newcomb's proposed law. He further pursued this discovery by collecting variety of data over the years. He finally published his observations of the first digit phenomenon over the wide variety of datasets.

The table below shows the predicted frequencies for first significant digits according to Benford's law.

First Significant digit	Frequency %
1	30%
2	18%
3	12%
4	10%
5	8%
6	7%
7	6%
8	5%
9	5%

Figure 5.1 Benford's Law standard values for significant digit

Figure 5.2 is the graph on the digit frequencies:

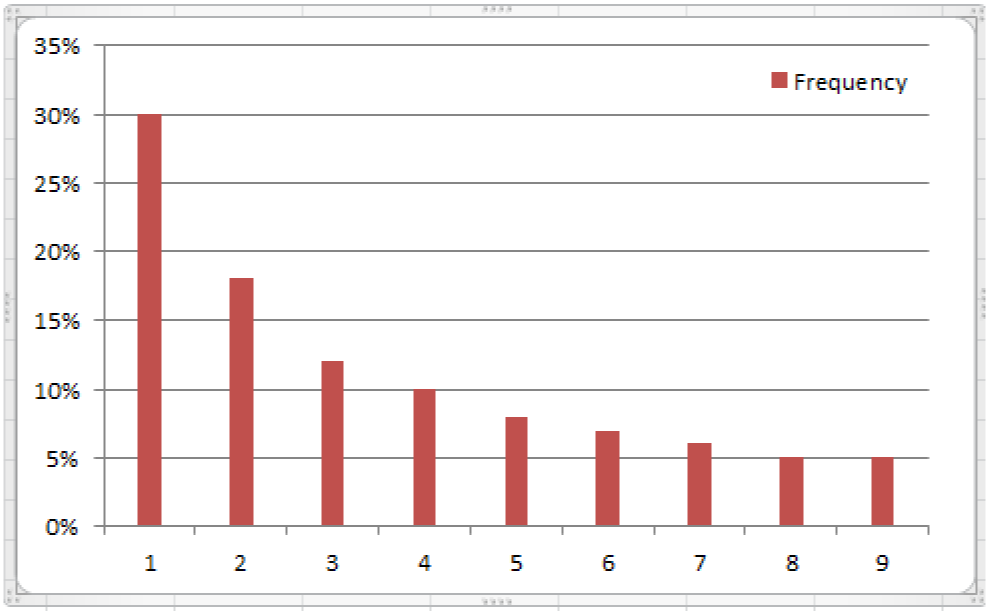


Figure 5.2 Graph plots of the percentage standard values of Benford's Law

Benford's Law can be generalized to higher digits as well. According to the second digit significant law the probability is given by:

Probability (second significant digit =d) = $\sum_{k=1}^{9}$ log $_{10}$ (1+(10k+d)$^{-1}$) .

This general law leads to a corollary that the significant digits are not independent on each other but have dependency. As we know the probability of the second digit being 2 will be approximately 0.109 according to the above formula. But according to the general law corollary, the probability that the second digit is 2, given the first digit is 1 is 0.115. This shows that there is dependency among the significance digits. The dependency is inversely proportional to distance between the significant digits in a uniform distribution. It decreases as the distance increases.

Benford's Law is the consequence of the sum of many independent random variables where the higher end of the range is different. It is rooted in the mixture of uniform distributions. The nodes in the social network are strongly correlated due to three reasons namely influence, homophily and environment [18]. Due to the strong correlation present in social network, the above laws of uniform distribution are not true for social network.

Benford's Law has proven to be helpful in detecting frauds in auditing and forensic accounting, since the fraudster is unlikely to be aware of it. It helps auditors in the planning phase of audit to identify the unusual transactions and also errors when used with transactional data. Auditors supervise and identify payment amounts for duplicates, invalid or missing check or invoice numbers. It is applied to an entire account and checked for invalid distribution of numbers and not just by sampling the account [14].

Although all the datasets obey Benford's first digit phenomenon, there are several other that do not [15]. Exceptions include the random numbers (like lottery), data with minimum and maximum constraints, artificially created data, dataset influenced by human psychology etc. The exceptions are:

- The value of the nodes should have same attribute.
- There should be no maximum or minimum value defined for that attribute.
- The values if numeric should not be pre-assigned. They should be random in nature.
- There should not be any human intervention in the data.

In 1994, Boyle showed that the data that obeys first digit phenomenon only when the data has random variables from different sources, being multiplied, divided or underwent any mathematical calculations [19]. Detection of fraud in any statistical data is merely manipulation by humans which turns to be behavioral intervention. This leads to an inference that only those datasets which are a result of statistics obey Benford’s law and those with human intervention do not obey.

4.2 Implementing Benford’s Law

Applying Benford’s First Digit Phenomenon on Facebook data of friend count of each of my friends in my friend list gives some surprising results. The steps to implement the law on Facebook data:

i. The count of friends of my friends is sorted in ascending order according to the first significant digit.

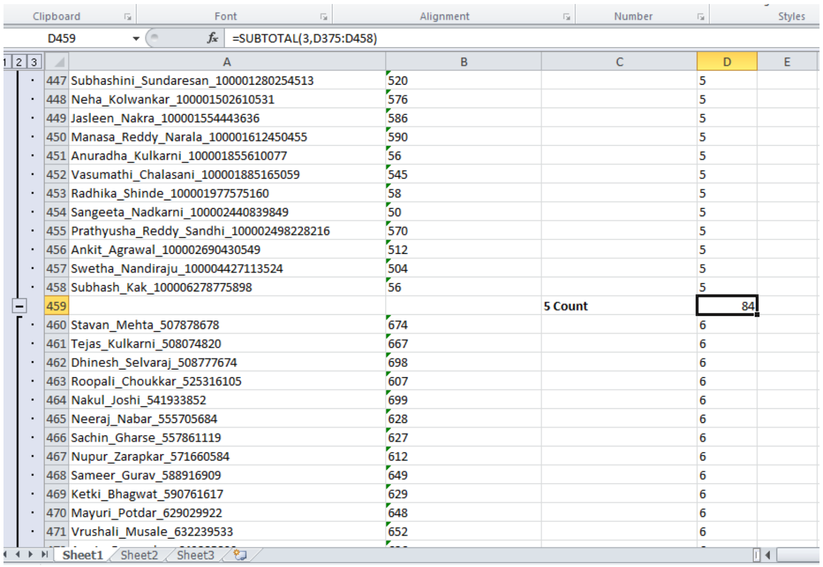
ii. The count of the groups formed is thereby noted.

iii. Table is formed with columns of significant digits from 1 to 9, the count of groups and the Benford’s standard values for respective significant digits.

iv. Graph is plotted with the above table formed.

v. The friend count and Benford’s values are in percentage unit.

The image below gives the idea how data was formatted to plot Benford’s law:

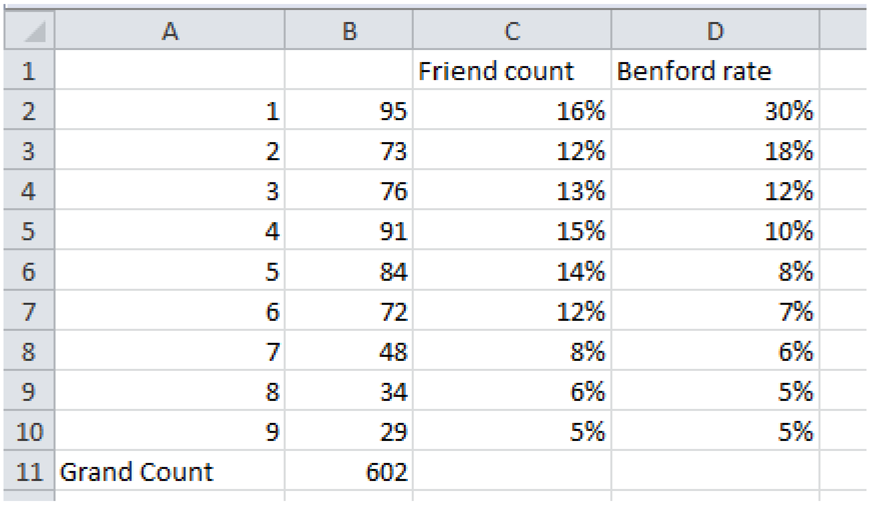


D459 =SUBTOTAL(3,D375:D458)

	A	B	C	D	E
447	Subhashini_Sundaresan_100001280254513	520		5	
448	Neha_Kolwankar_100001502610531	576		5	
449	Jasleen_Nakra_100001554443636	586		5	
450	Manasa_Reddy_Narala_100001612450455	590		5	
451	Anuradha_Kulkarni_100001855610077	56		5	
452	Vasumathi_Chalasani_100001885165059	545		5	
453	Radhika_Shinde_100001977575160	58		5	
454	Sangeeta_Nadkarni_100002440839849	50		5	
455	Prathyusha_Reddy_Sandhi_100002498228216	570		5	
456	Ankit_Agrawal_100002690430549	512		5	
457	Swetha_Nandiraju_100004427113524	504		5	
458	Subhash_Kak_100006278775898	56		5	
459			5 Count	84	
460	Stavan_Mehta_507878678	674		6	
461	Tejas_Kulkarni_508074820	667		6	
462	Dhinesh_Selvaraj_508777674	698		6	
463	Roopali_Choukkar_525316105	607		6	
464	Nakul_Joshi_541933852	699		6	
465	Neeraj_Nabar_555705684	628		6	
466	Sachin_Gharse_557861119	627		6	
467	Nupur_Zarapkar_571660584	612		6	
468	Sameer_Gurav_588916909	649		6	
469	Ketki_Bhagwat_590761617	629		6	
470	Mayuri_Potdar_629029922	648		6	
471	Vrushali_Musale_632239533	652		6	

Figure 5.3 Data formatting for Benford Law implementation

The friends data that was retrieved from Facebook is arranged is ascending order of the first digit of the friend count. Then groups are formed according to the first significant digit. The total count for each group is noted and used further for analysis. The final table that is generated:



	A	B	C	D
1			Friend count	Benford rate
2	1	95	16%	30%
3	2	73	12%	18%
4	3	76	13%	12%
5	4	91	15%	10%
6	5	84	14%	8%
7	6	72	12%	7%
8	7	48	8%	6%
9	8	34	6%	5%
10	9	29	5%	5%
11	Grand Count	602		

Figure 5.4 Final table for Benford law implementation

Figure 5.5 presents the results in a graphical form.

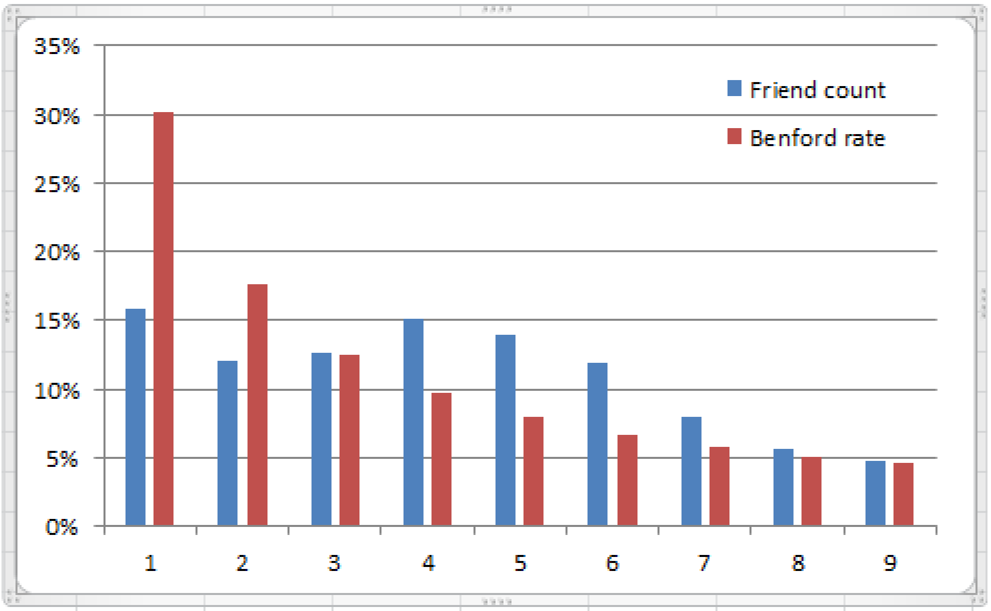


Figure 5.5 Benford’s Law implementation for Facebook data

As we can see from the above graph that friend count plotted does not obey Benford's Law of first digit phenomenon. The count of friends with significant digit as 1 is 15% and hence much less than that of standard Benford value i.e. 30%. It's the same for other values of friend count, either less than or greater than standard values. Concluding, Facebook data does not obey the law establishing that it has a behavioral component that alters the statistical frequencies.

Chapter VI

FURTHER ANALYSIS

The main objective of the thesis is to prove that social network paradoxes have a behavioral origin and not consequence of just mathematical properties of networks.

In Chapter III, we carried out statistical analysis on the Facebook data and the non-social data (road network) by calculating the mean and median of friends-of-friends and plotted the graph. As we saw for the road data, the mean equated or exceeded the median. For the Facebook data, the median exceeds the mean.

This characteristic of the data is a result of heterogeneous distribution of the degree of nodes in a social network. There are always some nodes with very high degree leading to formation of heavy tail in distribution which in turn skews the average.

Social network data like that of Facebook will always have heavy tail in the probability distribution skewed to the right (for larger degree nodes concentrated towards the left) due to the fact that friends have similar attributes, which is different from a non-social network where the nodes have independent attributes.

In the figure below we can see a highest degree node which is connected to maximum other nodes in the network. This node skews the average and results in large difference between its mean and median values giving rise to friendship paradox in.

mean. But for the Facebook data, it also exists in median which clearly indicates that it is not purely statistical but also has a behavioral origin

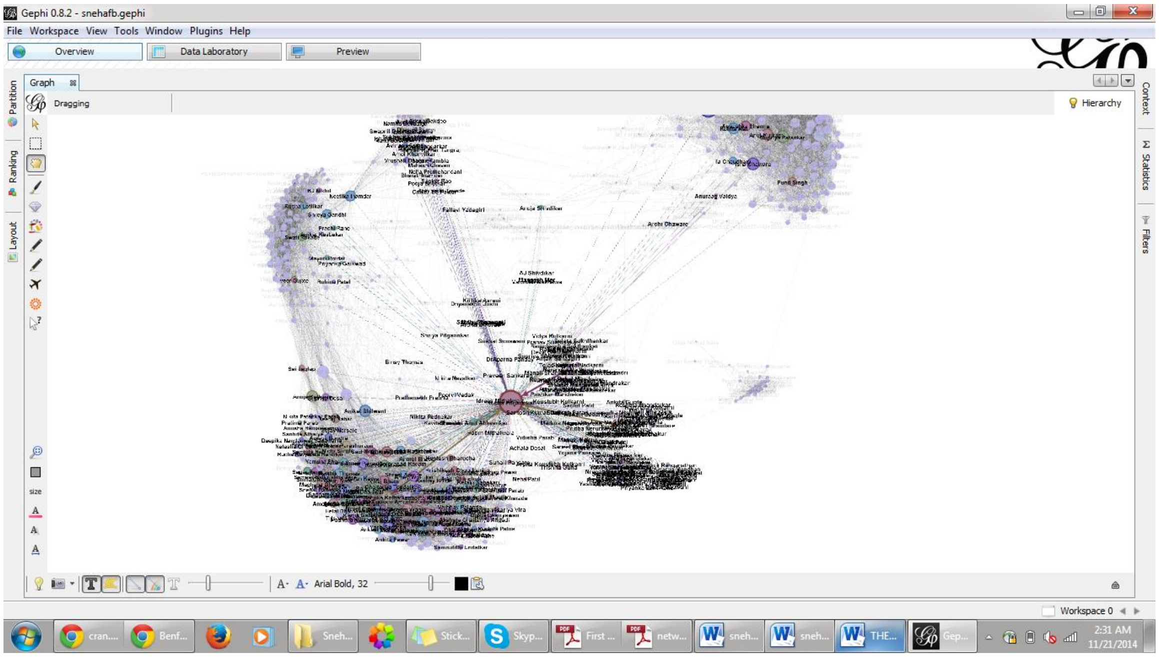


Figure 6.1 Gephi visualization of Facebook network showing the most popular node

Observe that the node attribute that we considered in the data analysis is the ‘friend count’. This attribute definitely changes from person to person in the sense whom to be added is completely the decision of the user. The user might add someone who shares a common community or who has an equal standard of living or he might add someone outside of his immediate interests. This decision is completely non-statistical in nature.

As discussed in an earlier chapter, a social network has two types of correlation: within-node-correlation and between-node-correlation. Within-node-correlation is a correlation existing between the user node and its degree while between-node-correlation is correlation between attributes of the user and those of its neighbors. Such correlations do not exist in networks governed completely by random behavior.

Social network characteristics are an aggregate of the behavior of many users who are driven by the behavior of their friends and with whom they might also be in competition. Thus social network paradoxes are not just result of mathematical results but have a behavioral component.

Chapter VII

CONCLUSION

The thesis has analyzed the friendship paradox of social networks. We did a statistical study of dataset retrieved from Facebook and we analyzed it for its relation between the mean and median of the distribution of friends of friends (FF). We found that in an overwhelming number of cases, the median (FF) was higher than the mean (FF). We also examined non-social network such as transportation network and found that in such cases the mean (FF) was greater than median (FF). This difference points to the behavioral basis of the social network paradox. In other words, the friendship paradox is not purely statistical in nature but has a behavioral component.

Further testing of the same data showed that it did not satisfy the Benford distribution. Since Benford's Law is not obeyed by psychology-driven data, our research establishes that friendship paradox related to mean and median of friends of friends has a behavioral origin.

REFERENCES

[1] R. Steinberg and W.I. Zangwill, (1983) *The prevalence of Braess's paradox*. Transportation Science, 17 (3) 301–318.

[2] G. Valiant and T. Roughgarden, (2006) *Braess' paradox in large random graphs*. Proceedings of the 7th Annual ACM Conference on Electronic Commerce (EC), 296–305.

[3] L. Rasumusson and M. Boman, (2002) *Analytical Expressions for Parrondo Games*, R71-R107; arXiv: physics/0210094v2.

[4] F. Kooti, N. O. Hodas and K. Lerman (2014) *Network Weirdness: Exploring the origins of Network Paradoxes* (CoRR) abs/1403.7242.

[5] H.-T. G. Chou, and N. Edge, (2012). *They are happier and having better lives than I am: the impact of using Facebook on perceptions of others' lives*. Cyerpsychology, Behavior, and Social Networking15 (2):117–121

[6] Pew Research Internet Project URl=http://www.pewinternet.org/2012/02/03/why-most-facebook-users-get-more-than-they-give

[7] N. Momeni, M. G. Rabbat (2014) *Measuring the Generalized Friendship Paradox in Networks with Quality-dependent Connectivity*; arXiv: 1411.0556 [cs.SI]

[8] Y.-H. Eom, H.-H. Jo (2014) *Tail-scope: Using friends to estimate heavy tails of degree distributions in large-scale complex networks* arXiv: 1411.6871v1 [physics.soc-ph]

[9] B. Fotouhi, N. Momeni, M. G. Rabbat (2014) *Generalized Friendship Paradox: An Analytical Approach*; arXiv: 1410.0586

[10] L. Weng, F. Menczer, and Y.-Y. Ahn, (2013) *Virality prediction and community structure in social networks*. Sci. Rep 3, 2522

[11] J. Lu (2013) *The Friendship Paradox* URL: http://web.williams.edu/Mathematics/sjmiller/public_html/hudson/The%20Friendship%20Paradox-%20Jackson%20Lu%204.pdf

[12] R. Kabacoff (2011), *R in Action*, Manning Publications Co. ISBN: 1935182390 9781935182399.

[13] A. Bogomolny, *Benford's Law and Zipf's Law*. URL: http://www.cut-the-knot.org/do_you_know/zipfLaw.shtml.

[14] M. Nigrini, (2000) *Digital Analysis Using Benford's Law: Tests Statistics for Auditors*. Vancouver, Canada: Global Audit Publications.

[15] EZ-R Stats, *Limitations of Benford's Law, and LLC Data Analysis made easier* URL: http://www.ezrstats.com/Benford6.htm

[16] T. P. Hill (1995) *A Statistical Derivation of Significant-Digit Law* Statistical Science Vol. 10

[17] F. Benford (1938). The law of anomalous numbers. Proceedings of the American Philsophical Society, 78, 551-572.

[18] A. Anagnostopou, R. Kumar, M. Mahdian (2008) *Influence and correlation in social networks*, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining ISBN: 978-1-60558-193-4

[19] C. Durtschi, W. Hillson and C. Pacini (2004). *The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data*; Journal of Forensic Accounting.

VITA

SNEHA PILGAONKAR

Candidate for the Degree of

Master of Science

Thesis: A STUDY OF NETWORK PARADOXES

Major Field: Computer Science

Biographical:

Education:

Completed the requirements for the Master of Science in Computer Science at Oklahoma State University, Stillwater, Oklahoma in December, 2014.

Completed the requirements for the Bachelor of Engineering in Computer Engineering at Mumbai University, Mumbai, Maharashtra/India in 2010.

Experience: Software Developer at Plant & Soil Science Laboratory, Agricultural Department, OSU from June 2013 to Present.