

Web Science cs532-s16

ASSIGNMENT 4 REPORT

DR. MICHAEL L. NELSON

BY: HUAN HUANG

02/25/2016

Problem 1

Determine if the friendship paradox holds for my Facebook account.* Compute the mean, standard deviation, and median of the number of friends that my friends have. Create a graph of the number of friends (y-axis) and the friends themselves, sorted by number of friends (x-axis). (The friends don't need to be labeled on the x-axis: just f1, f2, f3, ... fn.) Do include me in the graph and label me accordingly.

* = This used to be more interesting when you could more easily download your friend's friends data from Facebook. Facebook now requires each friend to approve this operation, effectively making it impossible.

I will email to the list the XML file that contains my Facebook friendship graph ca. Oct, 2013. The interesting part of the file looks like this (for 1 friend):

```
<node id="Johan_Bollen_1448621116">
  <data key="Label">Johan Bollen</data>
  <data key="uid"><![CDATA[1448621116]]></data>
  <data key="name"><![CDATA[Johan Bollen]]></data>
  <data key="mutual_friend_count"><![CDATA[37]]></data>
  <data key="friend_count"><![CDATA[420]]></data>
</node>
```

It is in GraphML format: <http://graphml.graphdrawing.org/>

Answer

For this problem, I had to parse the graphml file provided by Dr. Nelson. To do this, I received help from a fellow classmate Zhetan Li, who taught me how to use *pygraphml*. We had to tell the system to operate with UTF-8 coding since the file is coded with UTF-8 instead of ascii. Then we just use *pygraphml* as a parser and call for the needed attributes in each node.

```
1 from pygraphml import Graph
2 from pygraphml import GraphMLParser
3 import codecs
4 import sys
5
6 reload(sys)
7 sys.setdefaultencoding('utf-8')
8
9 savefile = open('faceFriendCount.txt', 'a')
10
11 parser = GraphMLParser()
```

```

12 g = parser.parse("mln.graphml")
13
14 nodes = g.nodes()
15 count = 0
16
17 for node in nodes:
18     count = count + 1
19     try:
20         friendName = node['name']
21         friendCount = node['friend_count']
22         print friendCount, friendName
23         savefile.write('{}{}'.format(friendCount, friendName))
24         savefile.write('\n')
25     except KeyError:
26         pass
27
28 mainUserName = "Michael L. Nelson"
29 print count, mainUserName
30 savefile.write('{}{}'.format(count, mainUserName))

```

To make the graph more comprehensive later, I sorted the list by numeric value. It is done in the terminal with the sort -n command, the output is saved in a file called sortedfacefriend.txt.

```

40 143,Fran Austin
41 144,Cynthia Suarez Eudy
42 147,Paul Ayris
43 155,Sally Jo Cunningham
44 165,Cathy Marshall
45 165,Michael L. Nelson
46 168,Julianne Allen
47 168,Lisa Owsley Rippy
48 170,Laura Garlow Yarborough
49 172,Brenda F. Nelson
50 181,Miriam Blake

```

Figure 1: Sample of part of sortedfacefriend.txt

Next, I load this sorted file into R and plotted the graph. The mean, median, and standard deviation are also calculated in R.

```

1 pdf("FaceFriends.pdf")
2
3 data2 = read.table(file.choose(), head=F, sep = ",")
4
5
6 plot(type="p", data2$V1, xlab = "Friends", ylab="Friend Counts",
7      main = "Facebook Friendship Paradox", xlim=c(0, 160),
8      ylim=c(0, 3500), las=0, col="red", pch=16, cex=0.3)
9 muser=which(data2$V2 == "Michael L. Nelson")
10 abline(v=muser, col=84, lty = 1)
11
12
13 legend(x=55,y=3400,c("Michael L. Nelson"),
14       col = 84, lty = 1)

```

```
15  
16 mean(data2$V1)  
17 #357.74  
18  
19 median(data2$V1)  
20 #259  
21  
22 sd(data2$V1)  
23 #370.7  
24  
25 dev.off()
```

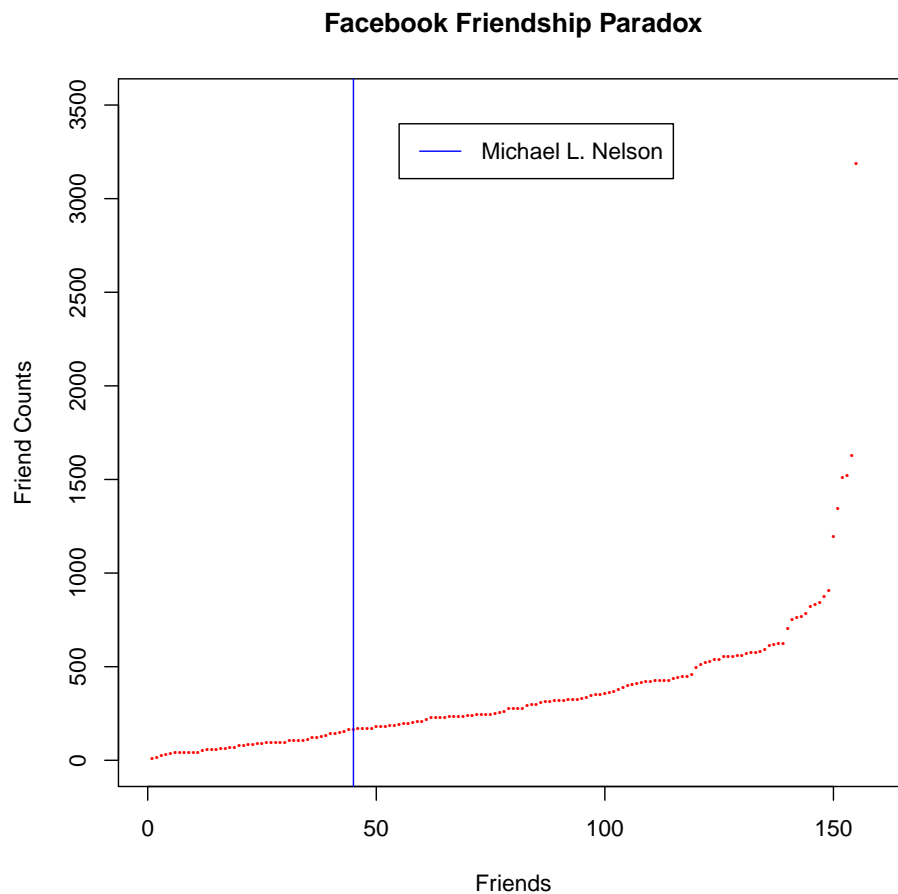


Figure 2: A dot graph of friend counts per friend

Dr. Nelson's Friends: 165

Mean: 357.74

Median: 259

Standard Deviation: 370.7

There are 11 individuals who do not have the “friend_count” attribute in their nodes. Therefore, they are not included in the graph and calculations.

The Friendship Paradox holds true for Dr. Nelson’s Facebook account. Looking at the graph, we can see that he is located at the left side of the graph, which means that most of his friends have more friends than him on Facebook.

Problem 2

Determine if the friendship paradox holds for your Twitter account. Since Twitter is a directed graph, use “followers” as value you measure (i.e., “do your followers have more followers than you?”).

Generate the same graph as in question #1, and calculate the same mean, standard deviation, and median values.

For the Twitter 1.1 API to help gather this data, see:

<https://dev.twitter.com/docs/api/1.1/get/followers/list>

If you do not have followers on Twitter (or don’t have more than 50), then use my twitter account “phonedude_mln”.

Answer

To solve this problem, I revisited a program from assignment 2 and made some changes. This time, instead of stream listen for links in people’s tweets, we are looking for the followers count of “phonedude_mln” and its followers. I decide to use tweepy this time, because it seems to be simpler and more suited to handle this problem. First, gain access with my consumer and access keys and tokens. Then, I get the follower count of “phonedude_mln” and write it to a file along with the screen name, which is “phonedude_mln”. Next, I get phonedude_mln’s followers and the follower count for each of them. I did encounter an issue at this part. To protect its server from overwhelming traffics, Twitter has set limitations on access rate. Since I used tweepy’s cursor function, I am limited to 15 calls in a period of 15 minutes. However, with 15 calls, I was only able to get 300 follower counts before being kicked out by Twitter(phonedude_mln had 489 followers at that time, the number has increased to 491 the next day). I did try to solve this issue by setting a sleeping period of 60s*15, but it gives me issues as well. At the end, I solved this problem by increasing the size of each call page. Tweepy cursor has a default setting of 20 items per call page, therefore, all I had to do is to increase the items per page. Just because I can, I set the page size to

its maximum, which is 200, and I was able to get all of the follower screen names and their follower counts in 3 calls.

```

1 #!/usr/bin/python
2 import time
3 import tweepy
4 from tweepy import OAuthHandler
5
6
7 savefile = open('followersCount.txt', 'a')
8
9 # load my API credentials
10
11 accesstoken = "110800099-ZiXf6imxd24OVqw5nnm4nkHXmVwhX46QQI0Wb3Zt"
12 accesstokensecret = "FIDFz2OI035FONaiL9uTzfqRauVArVwwVMuxiHDWQsmwo"
13 consumerkey = "1N9TkDHB9g4UIbjVj5j6idqnn"
14 consumersecret = "5Yjw1hRViWd5TloKLmEibPoregSltxO1bY7YJckjnecuSw4n9B"
15
16 # create twitter API object
17
18 auth = tweepy.OAuthHandler(consumerkey, consumersecret)
19 auth.set_access_token(accesstoken, accesstokensecret)
20
21
22 api = tweepy.API(auth)
23 mainuser = api.get_user('phonedude_mln')
24 userName = mainuser.screen_name
25 followersCounts = mainuser.followers_count
26 print followersCounts, userName
27 savefile.write('{:<10} {}'.format(followersCounts, userName))
28 savefile.write('\n')
29
30 users = tweepy.Cursor(api.followers, userName, count=200).items()
31
32 for follower in users:
33     try:
34         fcount = follower.followers_count
35         fname = follower.screen_name
36         print str(fcount) + " " + fname
37         savefile.write('{:<10} {}'.format(fcount, fname))
38         savefile.write('\n')
39     except tweepy.error.RateLimitError:
40         monitor_rate_limit=True, wait_on_rate_limit=True
41         continue
42     except StopIteration:
43         break

```

Again, I sorted the list by by using sort -n and saved to a file called sortedCount.txt.

308	468	tombortreasure
309	473	HistWebArchives
310	479	webrecorder_io
311	481	Philipp_Mayr
312	482	Jakdemir
313	489	phonedude_mln
314	498	bethcron
315	501	tsuomela
316	504	aag1091
317	505	mousta
318	505	NielsBr

Figure 3: Sample of part of sortedCount.txt

With this sorted file, I loaded the data into R and plotted the graph. Below are the R code and graphs.

```

1 pdf("TwitterFollowerslg.pdf")
2
3 data = read.table(file.choose(), head=F, sep = " ")
4
5
6 plot(type="p", data$V1, xlab = "Followers", ylab="Follower Counts",
7       main = "Twitter Friendship Paradox", xlim=c(0, 500),
8       ylim=c(1, 75000), las=0, col="red", pch=16, cex=0.3, log='y')
9 muser=which(data$V2 == "phonedude_mln")
10 abline(v=muser, col=84, lty = 1)
11 y = c(0, 10, 100, 1000, 10000, 100000)
12
13 legend(x=.1,y=70000,c("Michael L. Nelson"),
14        col = 84, lty = 1)
15
16 mean(data$V1)
17 #1045.867
18
19 median(data$V1)
20 #259
21
22 sd(data$V1)
23 #4146.195
24
25 dev.off()

```

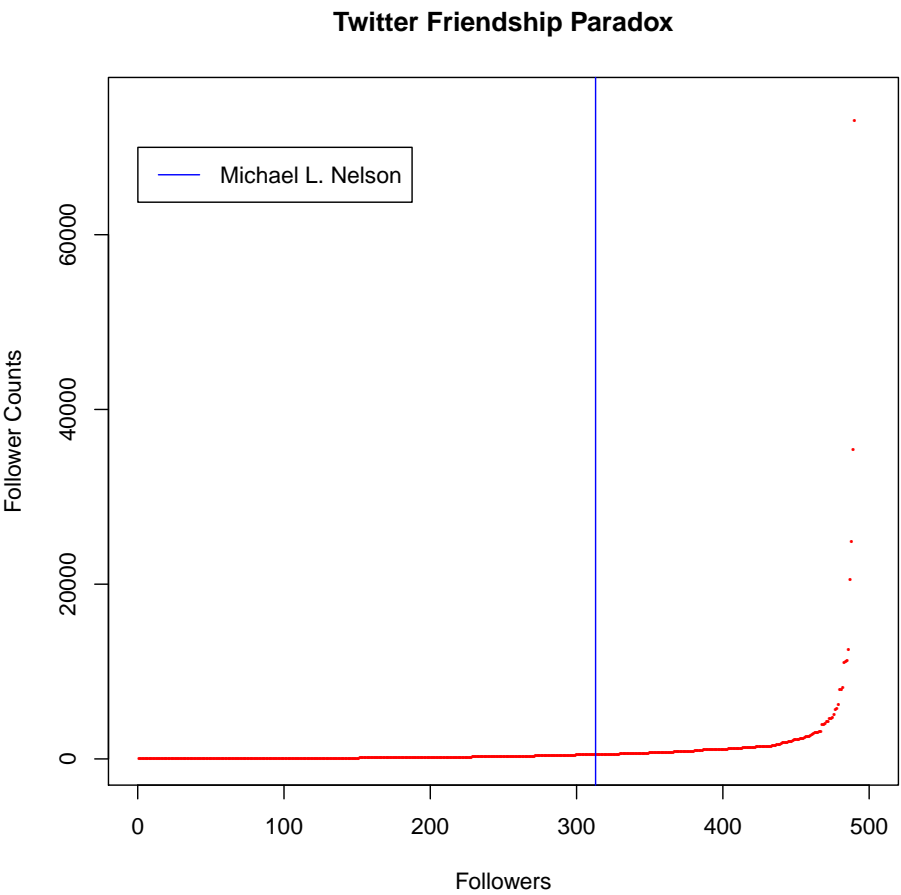


Figure 4: A dot graph of follower counts per follower

Because of few individuals who have many more followers than other users, it made this graph hard to read. To make this graph more comprehensive, I changed the y axis to log scale. However, logarithm does not work with 0, therefore, the six individuals who have 0 follower are not included in this graph.

1	0	ahmedkalka97
2	0	CrowNaito
3	0	hdo003
4	0	katjakra
5	0	unknown_6
6	0	witch_doctor81
7	1	amaranaas

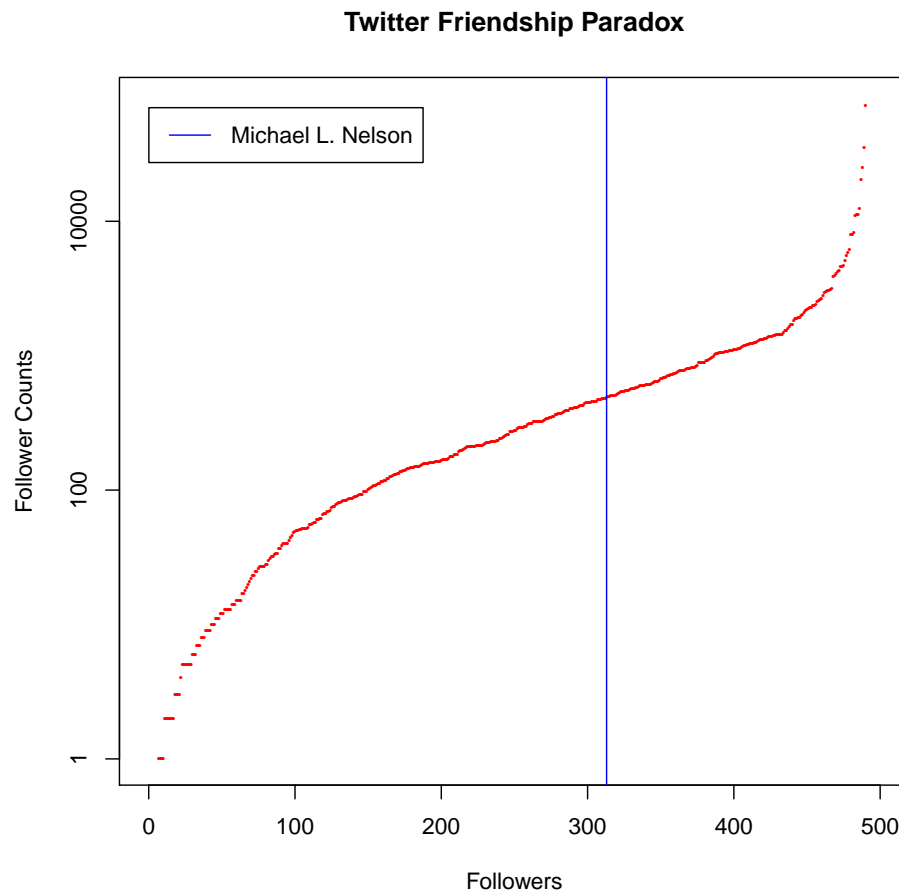


Figure 5: Second dot graph of follower counts per follower

Dr. Nelson's Followers: 489

Mean: 1045.867

Median: 259

Standard Deviation: 4146.195

Interesting enough, the median for Dr. Nelson's followers' follower counts is also 259. I thought that I made a mistake at first, but after checking several times, I am certain it is 259. However, the number has change now since Dr. Nelson has gained few more followers.

The Friendship Paradox does not hold true for Dr. Nelson's Twitter account. Based off the graph, we can see that Dr. Nelson is located towards the right side of the graph, which means that most of Dr. Nelson's followers have less followers than Dr. Nelson.