CS 432/532: Web Science

Spring 2017

Assignment 4

Michael Micros

Instructor: Michael L. Nelson

Old Dominon University
Norfolk, Virginia

March 2, 2017

Problem 1: Friendship Paradox for Dr. Nelson's Facebook

As part of the first problem, Dr. Nelson's facebook graphml file was provided. In order to extract the relevant information from this file the "pygraphml" library was imported. This library allows the user to parse the .graphml file and extract the nodes, isolating the desired information which in this count was the "friend_count" attribute. It was noted that not all of the users had a "friend_count" category. The friend count for all the friends was saved into a file and loaded to RStudio where the mean, median and standard deviation were calculated (performing the calculations in R was slightly more convenient, instead of importing "numpy" in Python). In the figure below 3 values stand out: 1) Dr. Nelson's friend count in Red, 2) The median in Green, 3) The mean in Yellow. Note that the Friendship Paradox is true, since Dr. Nelson's friend count is well below both the mean and the median.

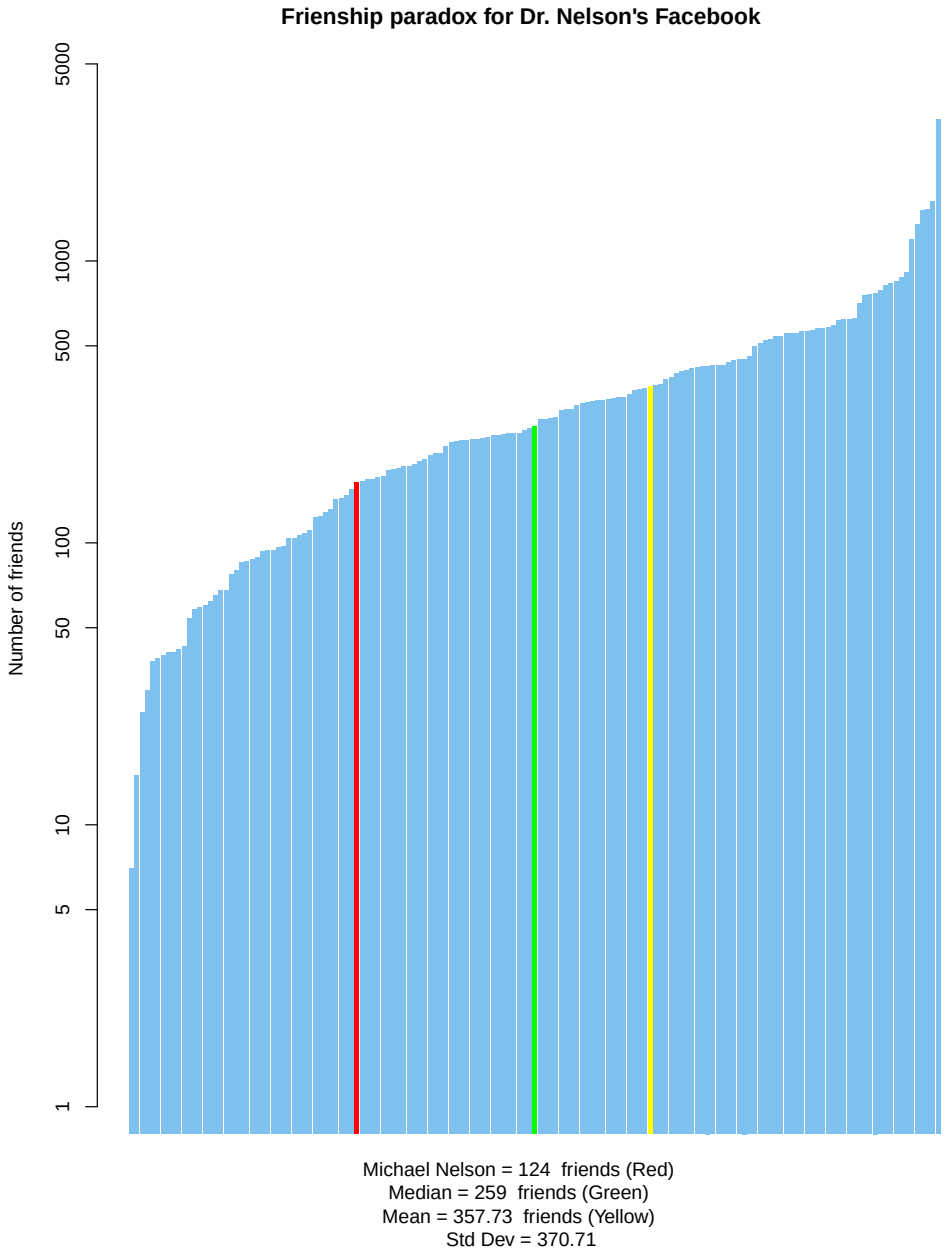

Figure 1: Plot of Dr. Nelsons friends VS their friend count

```
4   import pygraphml as p
5   import re
6   import numpy
7
8   def readFriends(filename):
9           f = open("graph",'w')
10
11          parser = p.GraphMLParser()
12          g = parser.parse(filename)
13
14          nodes = g.nodes()
15          print (len(nodes))
16          j=0
17          for node in nodes:
18                  f.write("Node "+ str(j)+"\n")
19                  j = j+1
20                  f.write(str(node)+"\n")
21
22
23  def getFriendCount(filename):
24          fin = open(filename,'r')
25          fout = open("Count",'w')
```

Figure 2: Part of the code in "p1.py" that parses the .graphml file and extracts the nodes

Problem 2: Friendship Paradox for Dr. Nelson's Twitter

Similarly to assignment 2, in order to gain access to data on Twitter, we must import tweepy and set up an API. Because I have only 2 followers, Dr. Nelson's twitter was used for this part of the assignment. First, the list of followers is collected and then the "followers_count" is isolated as can be seen in the figure below. Just as in Problem 1, the list was imported into RStudio for statistical calculations and plotting. We can note that once again the Friendship Paradox hold true, even though Dr. Nelson has more followers than half of his friends (median)

```
24  def getFollowers(api, user):
25          f = open("mlnFollowers",'w')
26          users = []
27          i=0
28          for page in tweepy.Cursor(api.followers, screen_name=user,count=200).pages():
29                  print(str(i))
30                  i=i+1
31                  users.extend(page)
32          for j in users:
33                  f.write(str(j.followers_count) + "\n")
34                  print(j.followers_count)
35
36          #comp = getStat(users)
```

Figure 3: "p2.py" Code

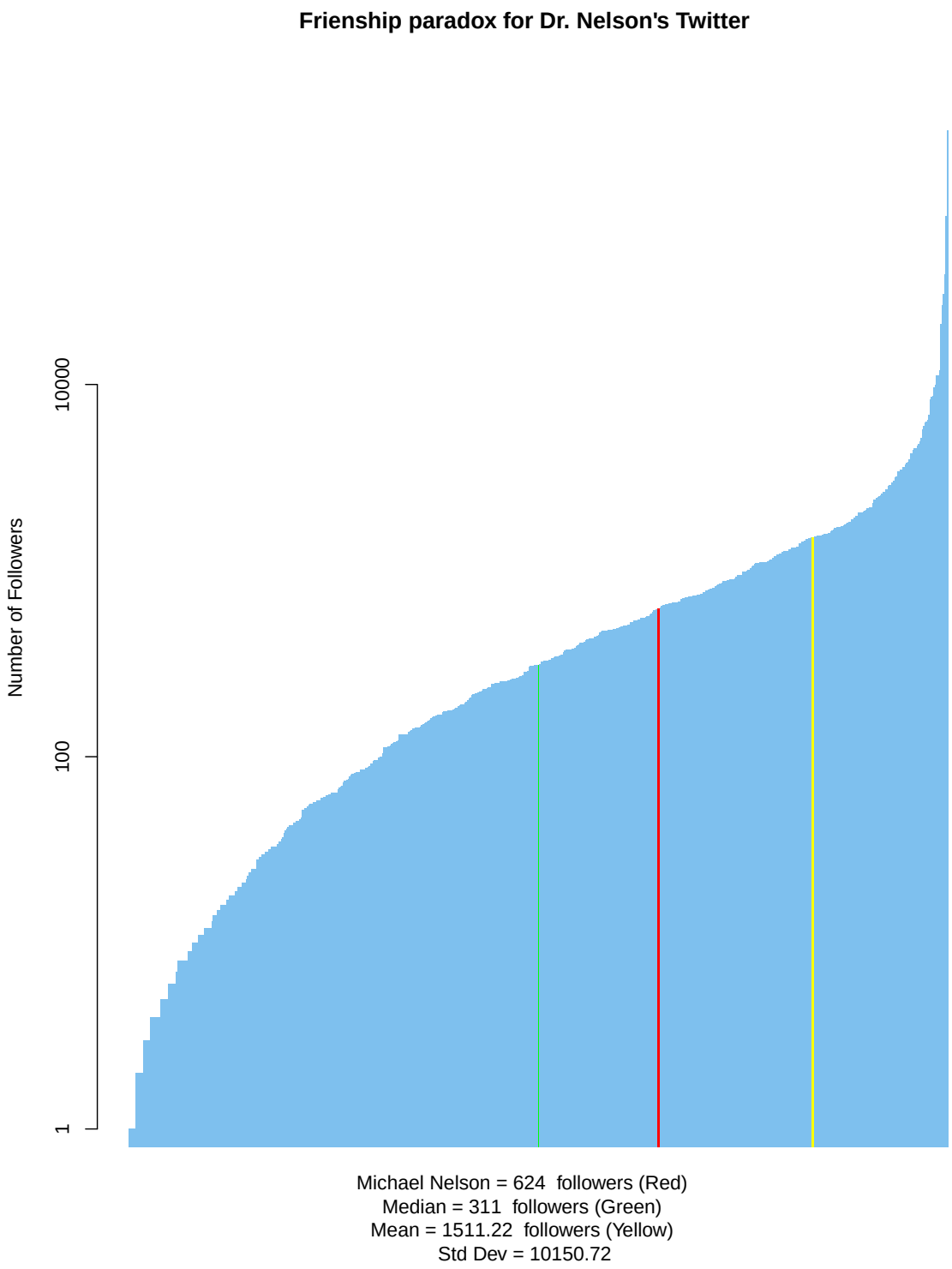

Figure 4: Plot of Dr. Nelsons followers VS their "followers_count"

Extra Credit:

I just wanted to let you know that I really appreciate and enjoy your class, and the fact that I have not done any of the extra credit offered in this and previous assignments is the result of terrible time management on my part and a heavy workload.