CS532 Web Science: Assignment 4

Finished on February 25, 2016

Dr. Michael L. Nelson

Naina Sai Tipparti
ntippart@cs.odu.edu

Contents

Listings

List of Figures

List of Tables

Problem 1

Question

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>
```

Listing 1: mln.graphml

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

Answer

A Python program, *qet_graphml.py*, has been written to extract number of friends that Dr. Nelson friends have. The program will search for this information in a file called *mln.graphml*. The output of this program would be like the following:

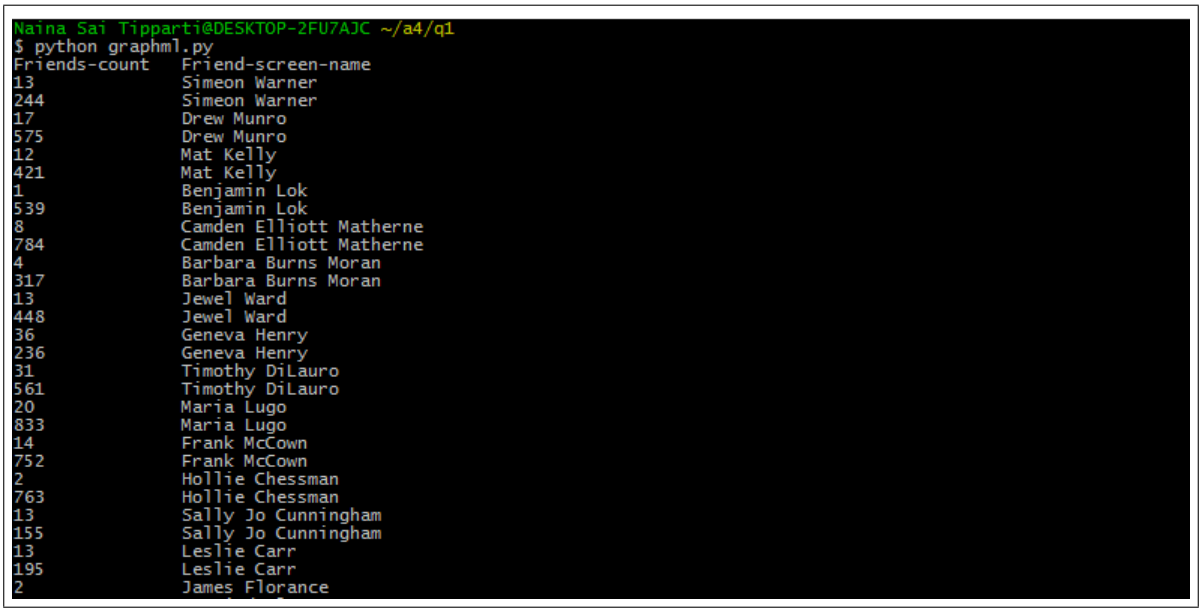

Figure 1: Sample output of number of friends

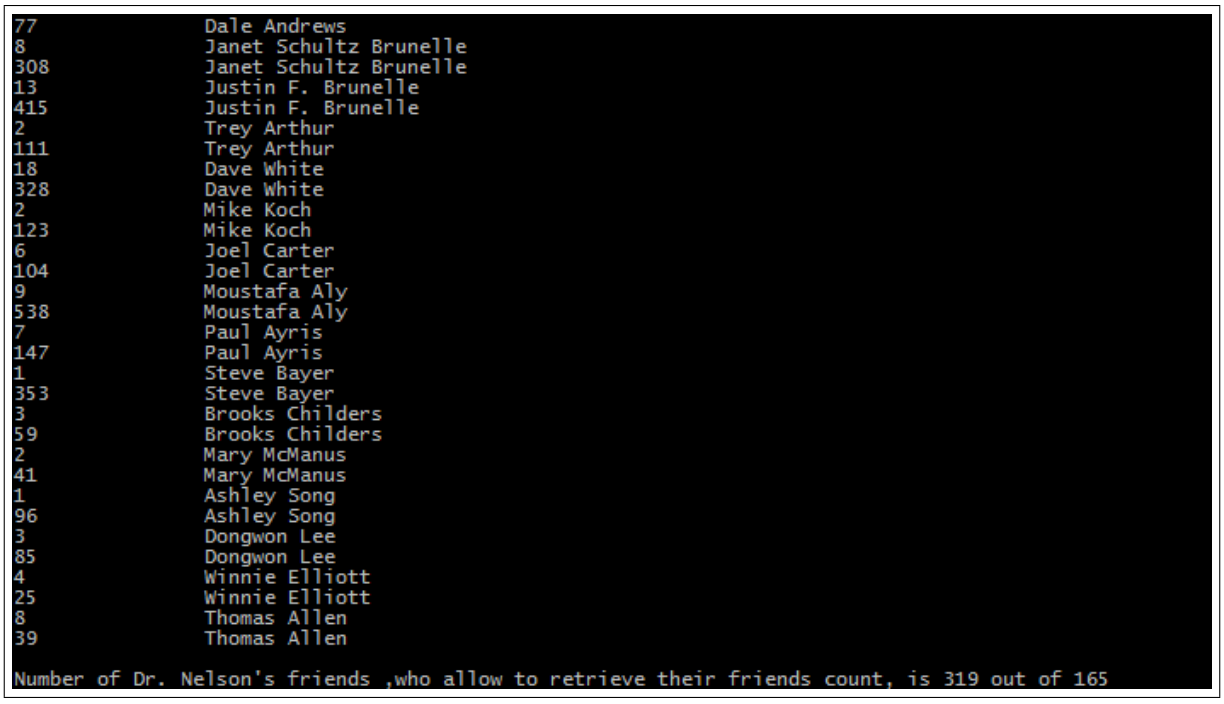

Figure 2: Sample output of number of friends

```
# -*- encoding: utf-8 -*-
#! /usr/bin/python
from __future__ import unicode_literals
import xml.etree.cElementTree as et
from bs4 import BeautifulSoup
from urlparse import parse_qs
import unicodedata
import urllib2
import re
import os
import sys

print '%-15s %-20s' %('Friends-count','Friend-screen-name')

file = "mln.graphml"
handler = open(file).read()
soup = BeautifulSoup(handler)
i = 0
all = 0
for message in soup.find_all('node'):
    all += 1
    foo = et.XML(str(message))
    name = ''
    for e in foo:
        if ('graphml_count' in str(e.items())):
            print '%-15s %-20s' %(e.text,name)
            with open('friend_counts', 'a') as outfile:
                outfile.write('%-15s %-20s\n' %(e.text,name))
            i += 1
        if ('name' in str(e.items())):
            name = e.text
print "\nNumber of Dr. Nelson's friends ,who allow to retrieve their friends count, is "+str
    (i)+" out of "+str(all)
```

Listing 2: get_graphml.py

I would like to let you know that even though Dr. Nelson have 319 friends, only 165 allow me to see their number of friends. This will affect the statistical result. For example, instead of dividing by 319 to get the mean, we divide by 165.

The `graphml_counts` file was ordered in place with the Unix command in Listing 3.

```
Naina Sai Tipparti@DESKTOP-2FU7AJC ~/a4/q1 cat graphml_counts | sort -g -o graphml_counts
```

Listing 3: Sort command

This file was then processed by the R script shown in Listing 4 to produce the graph in Figure 3

```
#! /usr/bin/Rscript

# read data
data <- read.table('D:/cs532/a4/q1/graphml_counts',sep=",")
x <- seq(1, length(data$V1))
y <- data$V1

# get notable values
mln_idx <- grep("phonedude_mln", data$V2)
med_val <- median(data$V1)
med_idx <- which(abs(y - med_val) == min(abs(y - med_val)))
mean_val <- mean(data$V1)
mean_idx <- which(abs(y - mean_val) == min(abs(y - mean_val)))
```

```
std_dev <- sd(data$V1)

# draw the graph
pdf("D:/cs532/a4/q1/facebook_graphml.pdf")
plot(x, y, type="l", log="y", pch=19, main="Dr. Nelson's Friends' Friends",
     ylab="Number of Friends", xlab="Index of Friend")

# illustrate points of interest
abline(h=data$V1[mln_idx], col="red")

# The Legend of the Data
legend(x=82, y=5, cex=0.8, lty=c(1, 1),
     col=c("red", "white", "white", "white", "white"),
     c(paste("Nelson: ", data$V1[mln_idx]), paste("median: ", med_val),
          paste("mean: ", format(round(mean_val, 4), nsmall = 4)),
          paste("std dev: ", format(round(std_dev, 4), nsmall = 4)),
          paste("median + 1 std dev: ", format(round(med_val + std_dev, 4), nsmall = 4))))
dev.off()
```

Listing 4: Graph Creation Script

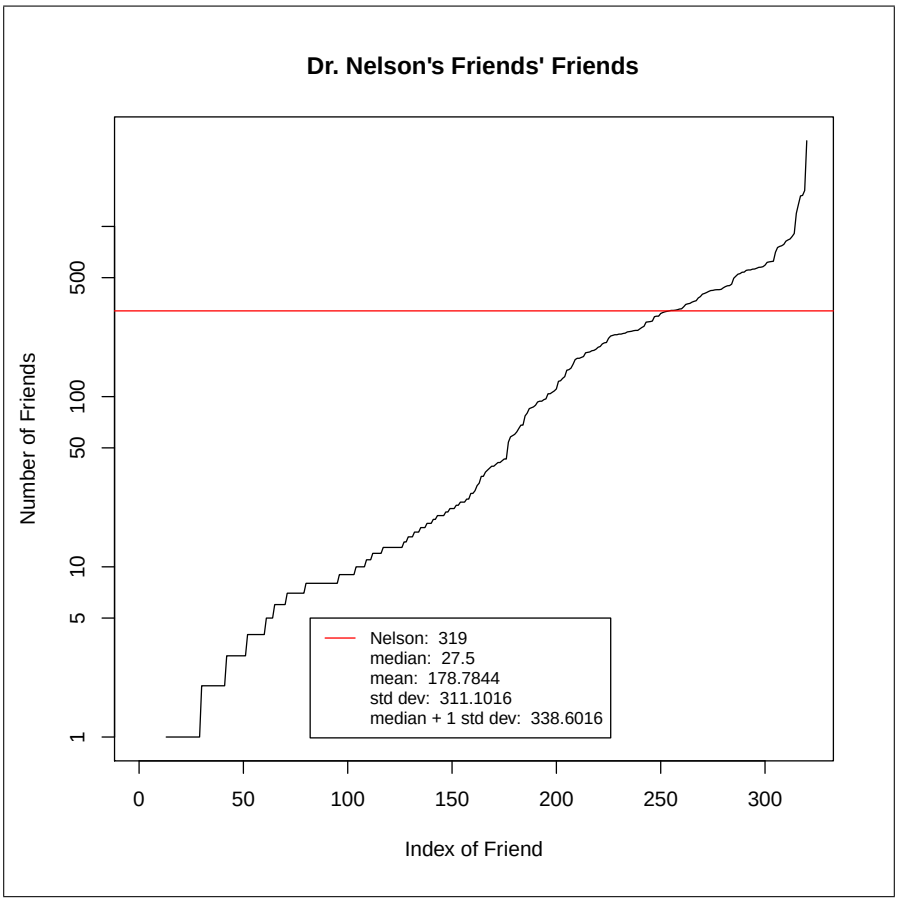

Figure 3: The Friendship Graph for Facebook

Mean	178.7844
Median	27.5
Std Dev	311.1016

Table 1: Statistics on the count of Dr. Nelson Facebook Friends' Friends, values straight from R

The median, mean and standard deviation were all calculated, with the median, mean and median plus one standard deviation.