CS595 Intro to Web Science, Assignment #5

Valentina Neblitt-Jones

October 17, 2013

The "friendship paradox" (`http://en.wikipedia.org/wiki/Friendship_paradox`) says that your friends have more friends than you do.

Question 1

Determine if the friendship paradox holds for your Facebook account. Create a graph of the number of friends (y-axis) and the friends sorted by number of friends (x-axis). (The friends don't need to be labeled on the x-axis.) Do include yourself in the graph and label yourself accordingly.

Compute the mean, standard deviation, and median of the number of friends that your friends have.

You can download your network in an XML file by using the NameGenWeb Facebook app:

`https://apps.facebook.com/namegenweb`

You will need to give this app permission to access your Facebook data. Make sure you select "Friend Count" as an Extended Attribute. When you download the data, download it in the GraphML format.

If you do not have a Facebook account, email me and I will send you my GraphML file.

Answer to Question 1

My GraphML file is available if necessary, but since it had real names in it I left it on my hard drive instead of loading it to GitHub. When I chose the anonymize option for generating the graph, it also did not provide the friend count so I had to run it without that option. The nodes in the file also contained a unique number for each person and in DoD we would say that it is sufficient for PII violation, but I could find no way to use the number to identify a person in Facebook. I did keep the names on a version of my program because I was curious about who my most popular friend was. Surprise! It was a computer scientist.

I used ElementTree to parse my XML with Python. There was a lot of misunderstanding as I tried to understand how to step through this XML to get at the elements I actually needed.

```
import xml.etree.ElementTree as ET

s = open('friendcountreport.csv', 'w')

tree = ET.parse('/Users/vneblitt/Documents/FacebookGraphs/ValentinaNeblitt-Jones_1381612295-ego.graphml')
root = tree.getroot()

namespace = '{http://graphml.graphdrawing.org/xmlns}'

# structure: graph -> node -> data

nodes = root.findall(namespace + 'graph/' + namespace + 'node')

myfriends = 0

s.write('Identifier' + ',' + 'Friend Count' + '\n')

for node in nodes:
    data = node.findall(namespace + 'data')
    friends = 0
    myfriends = myfriends + 1

    for datum in data:
        if datum.attrib.get('key') == 'uid':
            uid = datum.text
        if datum.attrib.get('key') == 'friend_count':
            friends = datum.text

    if friends != 0:
        s.write(str(uid) + ',' + str(friends) + '\n')

s.write('Valentina,' + str(myfriends) + '\n')
```

Figure 1: Showing Friend Count

I discovered when reviewing the GraphML file that some nodes did not contain a Friend Count. I investigated two of those cases and found a difference in summary friends data. Figure 2 shows a friend whose friend count was in the file and Figure 3 shows a friend whose friend count was not in the file. Some of my friends only allow me to see our mutual friends not all of their friends and therefore the application was unable to pull a friend count for those individuals. I have 230 friends, but my report on friend counts only has 208 of those friends' friend counts. This means that 22 of my friends have the mutual friends only setting. I omitted nodes that did not contain a friend count (Figure 4).

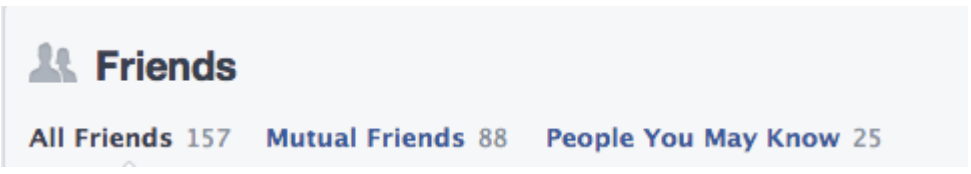

Figure 2: Showing Friend Count

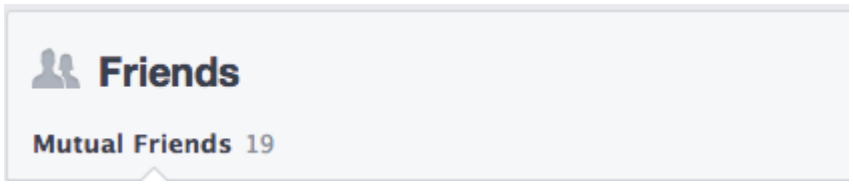

Figure 3: Hiding Friend Count

```
if friends != 0:
    s.write(str(uid) + ',' + str(friends) + '\n')
```

Figure 4: Code that Omits Friends with No Friend Count Available

Question 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 20), then use my Twitter account "phonedude_mln".

Answer to Question 2

Extra Credit - LinkedIn (2 points)

Repeat question #1, but with your LinkedIn profile.

Answer to Extra Credit - LinkedIn

Extra Credit - Twitter (1 point)

Repeat question #2, but change "followers" to "following"? In other words, are the people I am following following more people?

Answer to Extra Credit - Twitter

Resources

- Grosfield, Troy. Parsing XML with Python using ElementTree. http://blog.troygrosfield.com/2010/12/18/parsing-xml-with-python-using-elementtree/
- McCown, Frank. Producing Simple Graphs with R. http://www.harding.edu/fmccown/r/
- Poulson, Barton. R Statistics Essential Training. http://www.lynda.com/course20/R-tutorials/R-Statistics-Essential-Training/142447-2.html
- Python.org. The ElementTree XML API. http://docs.python.org/3.3/library/xml.etree.elementtree.html
- Seminar for Statistics. R Documentation: Arithmetic Mean. http://stat.ethz.ch/R-manual/R-devel/library/base/html/mean.html
- Seminar for Statistics. R Documentation: Concatenate Strings. http://stat.ethz.ch/R-manual/R-devel/library/base/html/paste.html
- Seminar for Statistics. R Documentation: Median Value. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/median.html
- Seminar for Statistics. R Documentation: Standard Deviation. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/sd.html
- Stack Overflow. Change colors or particular bars in a bar chart. http://stackoverflow.com/questions/13112974/change-colours-of-particular-bars-in-a-bar-chart
- Stack Overflow. How do you print to stderr in R? http://stackoverflow.com/questions/1109017/how-do-you-print-to-stderr-in-r
- Stack Overflow. Understanding the Order() function in R. http://stackoverflow.com/questions/2315601/understanding-the-order-function-in-r
- Twitter Developers Documentation. GET followers/list. https://dev.twitter.com/docs/api/1.1/get/followers/list