

Friendship Paradox

Web Science

10 points

The goal of this assignment is to write a Python script using the NetworkX package that examines the friendships of a given Facebook friendship network and determines if the friendship paradox is true.

The friendship paradox is that most of us have fewer friends than our friends do on average. The friendship paradox was first observed by the sociologist Scott L. Feld in 1991. It is a form of sampling bias in which people with greater numbers of friends have an increased likelihood of being observed among one's own friends. A 2012 study by Pew Internet¹ concluded the average Facebook user had 245 Facebook friends, but the average friend had 359 friends.

Data File

To perform this experiment, you will use data from Dr. McCown's Facebook social network circa 2013, which will be made available to you. The file is in GraphML format (<http://graphml.graphdrawing.org/>), a popular XML format for representing networks.

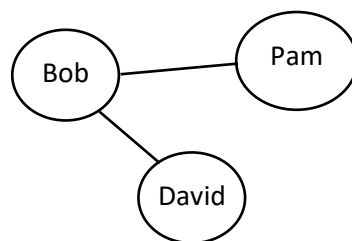
Each friend is a node in the network that has an entry like this:

```
<node id="Bob_White_5250237">
  <data key="Label">Bob White</data>
  <data key="uid"><![CDATA[5250237]]></data>
  <data key="name"><![CDATA[Joana Carlson]]></data>
  <data key="friend_count"><![CDATA[1261]]></data>
</node>
```

The friend_count data key contains the person's number of friends. Your mutual friends are represented as edges at the bottom of the file:

```
<edge id="3" source="Pam_Smith_54604435" target="Bob_White_5250237"></edge>
<edge id="4" source="David_Black_71001751" target="Bob_White_5250237"></edge>
```

The "source" is the person who initiated the friendship with the "target", but the graph is undirected, since Facebook friendships go both ways.



¹ Why most Facebook users get more than they give

<http://www.pewinternet.org/Reports/2012/Facebook-users/Summary/Friends-of-Friends.aspx>

NetworkX

Write a Python script that uses the NetworkX library to read the datafile:

```
import networkx as nx
G = nx.read_graphml('FrankMcCown-Facebook.graphml')
```

The following code snippet shows how to loop through the nodes of the graph and access the friend count:

```
for friend in G.nodes:
    # Make sure node has friend_count attribute
    if ('friend_count' in G.node[friend]):
        print(G.node[friend]['friend_count'])
```

Output

After analyzing the graph, your script needs to output the following information:

```
Total friends: 708
Average friend count: 627.3
How many have more friends: 206 (29.1%)
Most friends: Patti Jo Dillard White (3920)
Most friends in common: Becky 'Pratt' McCown (234)
Friendship paradox: YES or NO
```

Please use one decimal place in your answers, just like the example above.

- Line 1: How many friends are in the graph.
- Line 2: Average number of friends Dr. McCown's friends have.
- Line 3: Number of your friends (and percent) that have more friends than Dr. McCown.
- Line 4: The friend who has the most friends (friend count in parenthesis)
- Line 5: The friend that has the largest number of mutual friends (number of mutual friends in parenthesis).
- Line 6: Output YES if Dr. McCown is experiencing the friendship paradox or NO if not. Do not hard-code your output; make the output be based on the value of the related variables.

Submit

Submit your script to Canvas before it is due. If working in pairs, put both names in comments at the top of your script. When I run your script, I will use a different data file to produce different results, so don't hard-code anything about your program except the filename.