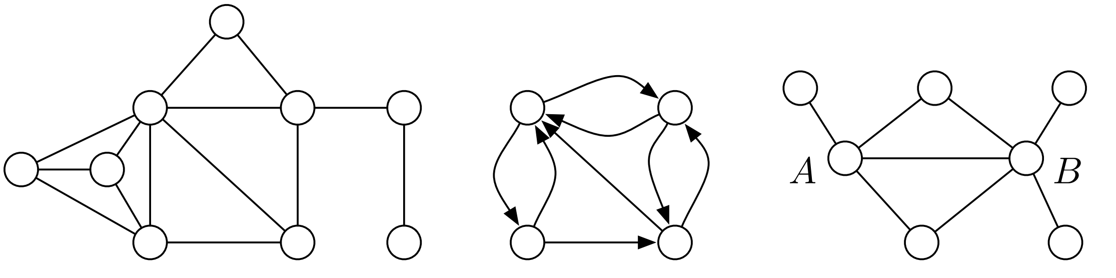
CSCI 5352 Network Analysis and Modeling Fall 2020

Prof. Daniel Larremore

Problem Set 2

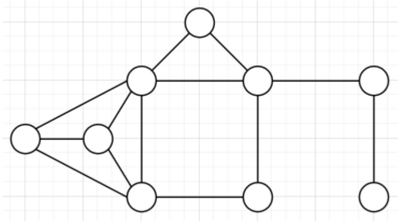
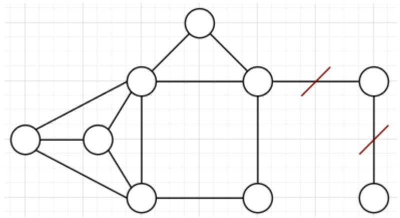
Student: Alex Book; Collaborators: Cole Sturza, Members of HW Group

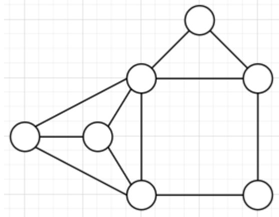
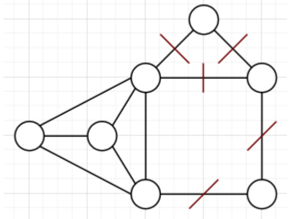
1. (13 pts) Consider the following three networks:

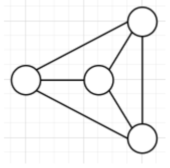


(A) (B) (C)

- (4 pts) Find a 3-core in network (A).
  The following images are the steps taken to find a 3-core:

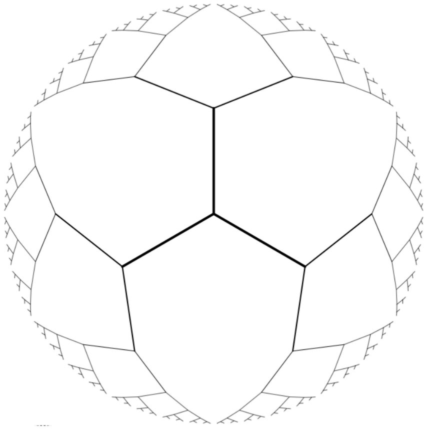
1-core:  transition from 1-core to 2-core: 

2-core:  transition from 2-core to 3-core: 

3-core: 

- (5 pts) What is the reciprocity of network (B)?
  Total edges = 8
  Reciprocated edges = 6
  Reciprocity = Reciprocated edges/Total edges = 6/8 = .75
- (4 pts) What is the cosine similarity of vertices A and B in network (C)?
  $\sigma_{AB} = \frac{n_{ij}}{\sqrt{k_A k_B}} = \frac{2}{\sqrt{4*5}} = \frac{1}{\sqrt{5}}$

2. (15 pts) A Cayley tree is a symmetric regular tree in which each vertex is connected to the same number k of others, until we get out to the leaves, like the figure below, with $k = 3$. Show that the number of vertices reachable in d steps from the central vertex is $k(k-1)^{d-1}$ for $d \geq 1$. Then give an expression for the diameter of the network in terms of k and the number of vertices n. State whether this network displays the "small-world effect," defined as having a diameter that increases as $O(\log n)$ or slower.



Proof by induction:

Base Case:
Let $d = 1$. In 1 step, there would be k nodes reachable from the central vertex (the k nodes it is connected to).

$$k(k-1)^{d-1} = k(k-1)^{1-1} = k \tag{1}$$

Let $d = 2$. In 2 steps, there would be $k(k-1)$ nodes reachable from the central vertex (the k nodes it is directly connected to, plus the $k-1$ nodes that each of *those* are connected to).

$$k(k-1)^{d-1} = k(k-1)^{2-1} = k(k-1) \tag{2}$$

Inductive Hypothesis:
Assume that for $d = i$, where $i \geq 1$, the number of nodes reachable from the central vertex is represented by $s_i = k(k-1)^{d-1}$.

Inductive Step:
Let $d = i + 1$. We can say that the number of nodes reachable s from the central vertex in any given step is equal to the number reachable from the last step multiplied by the quantity $(k-1)$ (as $k-1$ new nodes are reached from each existing 'leaf' node).

$$\begin{aligned} s_{i+1} &= (k-1) * s_i \\ s_{i+1} &= (k-1) * k(k-1)^{i-1} && \text{by inductive hypothesis} \\ s_{i+1} &= k(k-1)^i && \text{by multiplication} \end{aligned}$$

The diameter of a Cayley tree can be expressed as twice the number of steps taken d, as one edge is added on either 'side' of the tree with each step.
The total number of nodes n can be expressed as:

$$n = 1 + k\sum_{i=1}^{d}(k-1)^{i-1} \tag{3}$$

$$\text{with } j = i - 1\text{: } n = 1 + k\sum_{j=0}^{d}(k-1)^{j} \tag{4}$$

$$n = 1 + k\frac{1-(k-1)^{d+1}}{1-(k-1)} \tag{5}$$

$$n = \frac{2-k(k-1)^{d+1}}{2-k} \tag{6}$$

$$n(2-k) = 2 - k(k-1)^{d+1} \tag{7}$$

$$2 - n(2-k) = k(k-1)^{d+1} \tag{8}$$

$$\frac{2-n(2-k)}{k(k-1)} = (k-1)^{d} \tag{9}$$

$$\log(\frac{2-n(2-k)}{k(k-1)}) = \log((k-1)^{d}) \tag{10}$$

$$d = \frac{\log(\frac{2-n(2-k)}{k(k-1)})}{\log(k-1)} \tag{11}$$

$$\text{diameter} = 2d = 2\frac{\log(\frac{2-n(2-k)}{k(k-1)})}{\log(k-1)} \tag{12}$$

$$\text{diameter} \sim O(log(n)) \tag{13}$$

Because this network's diameter grows at a rate equivalent to $O(log(n))$, it does indeed display the "small-world effect." Note that the above holds true only for $k \geq 3$. Thus we can say that for $k \geq 3$, Cayley Trees display the "small-world effect."

3. (35 pts total) In this question, we will investigate several properties of online social networks by analyzing the Facebook100 ("FB100") data set, which you may download via the link posted on Piazza. Each of the 100 plaintext ASCII files in the FB100 folder contains an edge list for a 2005 snapshot of a Facebook social network among university students and faculty within some university. Interpret this edge list as a simple graph.[1]

   (a) (5 pts) In most social networks, we observe a surprising phenomenon called the *friendship paradox*. Let k_u denote the degree of some individual u, and let some edge $(u, v) \in E$. The paradox is that the average degree of the neighbor $\langle k_v \rangle$ is *greater* than the average degree $\langle k_u \rangle$ of the vertex. That is, on average, each friend of yours has more friends than you.

   The mean neighbor degree (MND) of a network is defined as

$$\langle k_v \rangle = \frac{1}{2m} \sum_{u=1}^{n} \sum_{v=1}^{n} k_v A_{uv} \quad . \tag{14}$$

   Derive an expression for $\langle k_v \rangle$ in terms of the average squared-degree $\langle k^2 \rangle$ and the average degree $\langle k \rangle$. Show your work.

   Note: In an undirected graph (which Facebook is, as friendship must be accepted/reciprocated), the adjacency matrix is symmetrical across the diagonal, so $A_{ij} = A_{ji}$.

$$\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{2m}{n} \tag{15}$$

$$\langle k^2 \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i^2 \tag{16}$$

$$k_i = \sum_{j=1}^{n} A_{ij} = \sum_{j=1}^{n} A_{ji} \tag{17}$$

$$\langle k_v \rangle = \frac{1}{2m} \sum_{u=1}^{n} \sum_{v=1}^{n} k_v A_{uv} \tag{18}$$

$$= \frac{1}{2m} \sum_{v=1}^{n} k_v \sum_{u=1}^{n} A_{uv} \tag{19}$$

$$= \frac{1}{2m} \sum_{v=1}^{n} k_v^2 \quad \text{by (18)} \tag{20}$$

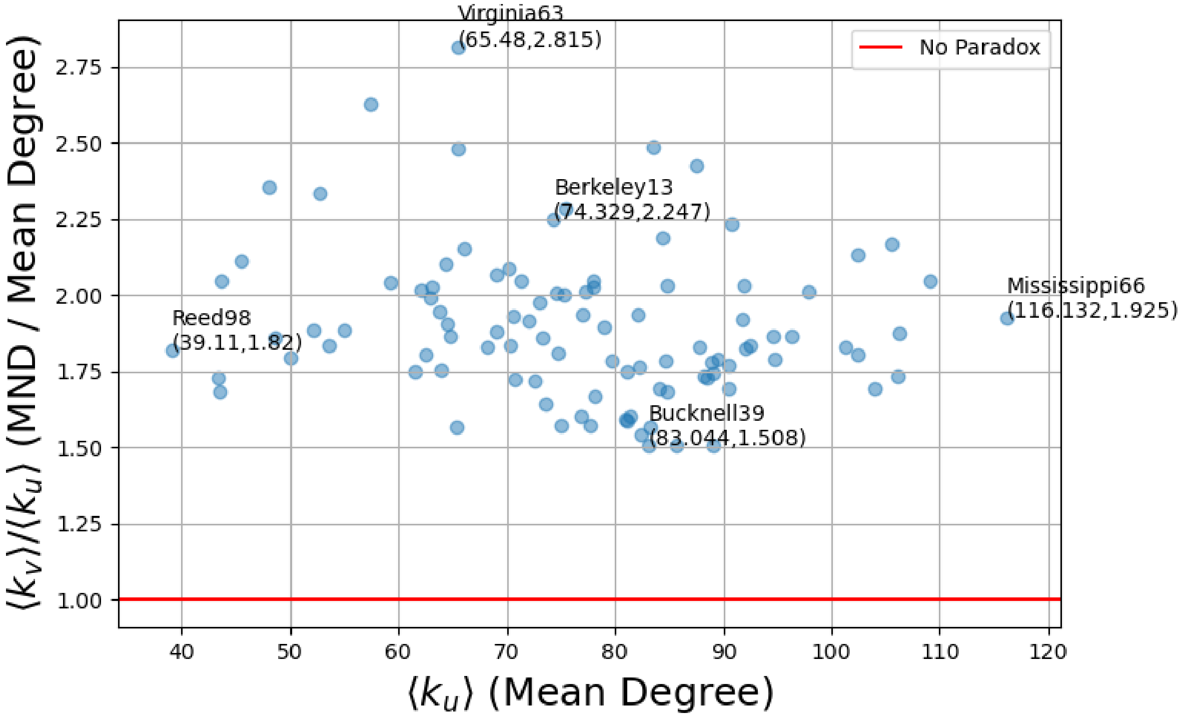
$$= \frac{1}{\langle k \rangle n} \sum_{v=1}^{n} k_v^2 \tag{21}$$

$$= \frac{\langle k^2 \rangle}{\langle k \rangle} \tag{22}$$

[1] The data were kindly provided by A.L. Traud, P.J. Mucha and M.A. Porter, as part of their paper "Social Structure of Facebook Networks," *Physica A* **391**, 4165–4180 (2012), which is freely available at `http://arxiv.org/abs/1102.2166` or `http://bit.ly/1ztbVoS`.

(b) (15 pts) Now, using all 100 of the FB100 networks, make a figure showing a scatterplot of the ratio $\langle k_v \rangle / \langle k_u \rangle$ as a function of the mean degree $\langle k_u \rangle$. Include a horizontal line representing the line of "no paradox," and label the nodes corresponding to Reed, Bucknell, Mississippi, Virginia, and UC Berkeley. (Remember: figures without axes labels will receive no credit.)

Comment on the degree to which we do or do not observe a friendship paradox across these networks as a group. Comment on whether there is any dependency between the magnitude of the paradox (the size of the MND, relative to the size of the mean degree) and the network's mean degree. A few points of extra credit will be awarded to an explanation of why we should, in fact, expect to see a friendship paradox in these networks, and that identifies the conditions under which we should expect to see *no* paradox.



There does indeed seem to be a friendship paradox across all schools, as can be seen in the figure generated by my code (figure shown above, code shown at the end of this document). There doesn't seem to be any direct relationship between the magnitude of the paradox and the network's mean degree.

We should expect to see a friendship paradox in these networks due to the right skew the mean degree has (few outliers that have a high degree, with many nodes having a low degree). This would cause a high likelihood of low-degree nodes to be connected to a high-degree node, disproportionately affecting its mean neighbor degree. For example, if a node a has degree 3, and its neighbors have degrees of 2, 2, and 50, the mean neighbor degree of a (18) is very heavily impacted by the one high-degree node it is connected to, despite the majority of a's neighbors having lower degrees than a itself.

The conditions under which we should expect to see no paradox would be that of a more uniformly-distributed mean degree (i.e. ring networks), where most nodes are similar or identical in degree, leading to very few nodes having a mean neighbor degree greater than their own degree.

(c) (15 pts) A related phenomenon in social networks is the *majority illusion*. Let $x \in \{0, 1\}$ be a binary-valued vertex-level property, and let $q = \frac{1}{n}\sum_u x_u$ be the fraction of vertices that exhibit this property. If we set $q < 0.5$, then this property appears only in a minority of nodes. The majority illusion occurs when $q < 0.5$, but the majority of a node's neighbors, on average, exhibit that property, that is, $\langle x_v \rangle > 0.5$.
Explain in words and mathematics how this can be possible.

This is due to many of the nodes that exhibit the property being well-connected (having a high degree). On average, a given node likely wouldn't have the property, but would likely be connected to nodes with the property (since the well-connected nodes have so many connections, including to many poorly-connected nodes). For an explanation, replace k_v with x_v in the equation for the mean neighbor degree (MND):

$$\langle k_v \rangle = \frac{1}{2m}\sum_{u=1}^{n}\sum_{v=1}^{n} k_v A_{uv} \tag{23}$$

$$\langle x_v \rangle = \frac{1}{2m}\sum_{u=1}^{n}\sum_{v=1}^{n} x_v A_{uv} \tag{24}$$

$$\langle x_v \rangle = \frac{1}{2m}\sum_{v=1}^{n} x_v \sum_{u=1}^{n} A_{uv} \tag{25}$$

$$\langle x_v \rangle = \frac{1}{2m}\sum_{v=1}^{n} x_v k_v \tag{26}$$

This shows that the mean neighbor property degree $\langle x_v \rangle$ increases more with neighbors that both have the property ($x_v = 1$) and have a high degree (k_v is relatively large). The majority illusion is explained by well-connected nodes *with* the property 'making up for' the many more less-connected nodes *without* the property.

(d) (20 pts *extra credit*) Another common property of social networks is that they have very small diameters relative to their total size. This property is sometimes called the "small-world phenomenon" and is the origin of the popular phrase "six degrees of separation".[2]

[2] This term originated in a play written by John Guare in 1990, which was turned into a 1993 movie starring Will Smith. The concept, however, was originated by Stanley Milgram, working in 1967, who was the first to measure the lengths of paths in large social networks. This is the same Stanley Milgram who did important but problematic work on obedience. Dive into this story at the Wikipedia page for "Milgram experiment" when you have time...

- For each FB100 network, compute (i) the diameter $\ell_{\max}$ of the largest component of the network and (ii) the mean geodesic distance $\langle \ell \rangle$ between pairs of vertices in the largest component of the network. Make two figures, one showing $\ell_{\max}$ versus network size n and one showing $\langle \ell \rangle$ versus the size of the largest component n. Comment on the degree to which these figures support the six-degrees of separation idea.
- Briefly discuss whether and why you think the diameter of Facebook has increased, stayed the same, or decreased relative to these values, since 2005. (Recall that Facebook now claims to have roughly 10^9 accounts.)

Code for 3b:

```
import networkx as nx
from os import listdir
from pprint import pprint
import matplotlib
from matplotlib import pyplot as plt
import pandas as pd

def fillDictFromFiles(fileList):
    dict = {}

    for file in filepaths:
        G = nx.read_edgelist(file)

        # finds the mean degree and mean neighbor degree of each network
        totalDegree = 0
        totalDegreeSquared = 0
        for i in G.degree:
            totalDegree += i[1]
            totalDegreeSquared += i[1]**2
        averageDegree = totalDegree/len(G)
        averageNeighborDegree = (totalDegreeSquared/len(G))/averageDegree

        # neighborDict = nx.average_neighbor_degree(G)
        # totalNeighborDegree = 0
        # for key in neighborDict:
        #     totalNeighborDegree += neighborDict[key]
        # averageNeighborDegree = totalNeighborDegree/len(neighborDict)

        school = file.split(’/’)[1].split(’.’)[0]
        print(school)

        dict[school] = (averageDegree, averageNeighborDegree)

    return dict

def plot_csv(file):
    df = pd.read_csv(file)
    df.columns = [’School’, ’Mean Degree’, ’Mean Neighbor Degree’]
    df[’Ratio’] = df[’Mean Neighbor Degree’]/df[’Mean Degree’]

    fig, ax = plt.subplots()
    ax.scatter(df[’Mean Degree’], df[’Ratio’], alpha = .5)
    ax.hlines(1, min(df[’Mean Degree’])-5, max(df[’Mean Degree’])+5, color=’red’,
        label=’No Paradox’)
    ax.grid(True)
    ax.set_xlim(min(df[’Mean Degree’])-5, max(df[’Mean Degree’])+5)
    ax.set_xlabel(r’$\langle k_{u}\rangle$ (Mean Degree)’, fontsize=18)
    ax.set_ylabel(r’$\langle k_{v}\rangle / \langle k_{u}\rangle$ (MND / Mean Degree)’,
        fontsize=18)

    # only label the desired schools
    labeldf = df.loc[df[’School’].isin([’Reed98’, ’Bucknell39’, ’Mississippi66’,
                                        ’Virginia63’, ’Berkeley13’])]
    for i in labeldf.index:
        ax.annotate(df[’School’][i] + ’\n(’ + str(round(df[’Mean Degree’][i],3)) + ’,’
                                            + str(round(df[’Ratio’][i],3)) + ’)’,
                    (df[’Mean Degree’][i], df[’Ratio’][i]))

    ax.legend()
    plt.show()

if __name__ == ’__main__’:
    path = ’facebook100txt’

    filepaths = []

    for f in listdir(path):
        # only select the desired files
        if (’_attr’ not in f) and (’Traud’ not in f) and (’readme’ not in f):
            filepath = path + ’/’ + f
            filepaths.append(filepath)

    dict = fillDictFromFiles(filepaths)
    df = pd.DataFrame.from_dict(data=dict, orient=’index’)
    df.to_csv(’dict_file.csv’, header=[’Average Degree’, ’Average Neighbor Degree’])

    plot_csv(’dict_file.csv’)
```