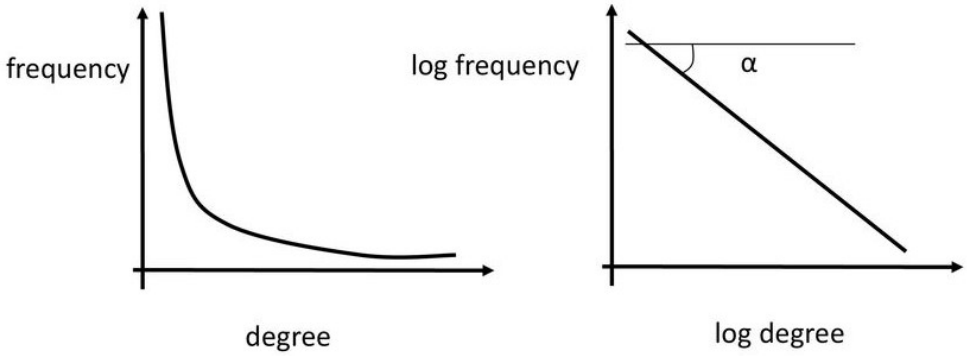
Project 3: Network Properties in Spark

CSC 591: Algorithms for Data-Guided Business Intelligence
Rachit Shah (rshah25@ncsu.edu)
February 13, 2019

Problem 1: Degree Distribution

In this problem we have to compare whether the distribution of degrees obtained from the graphs are scale free or not, and hence essentially finding whether they follow the power law or not. A distribution that follows power law looks like the plot given below. The log-log plot of the distribution is linear and can be fitted with a straight line. Hence, we will use np.polyfit() function with degree 1 to find the fit for the log distribution and see whether it's a good fit.

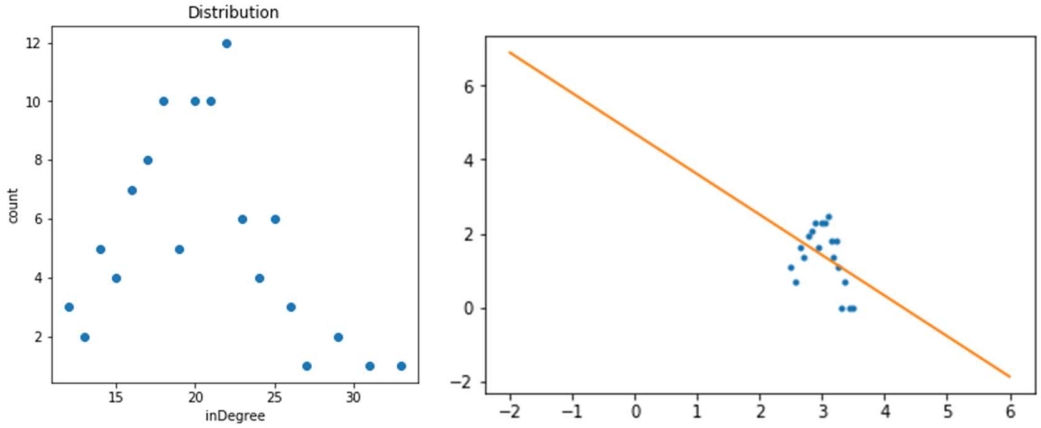
The script to generate mentioned plots and values can be found as a jupyter notebook with the name (power lab.ipynb). You'll need to run a jupyter notebook in a Python3 environment and pass the name of the csv file to assess in the filename variable in the notebook.



1. **Generate a few random graphs. You can do this using networkx's random graph generators. Do the random graphs you tested appear to be scale free? (Include degree distribution with your answer).**

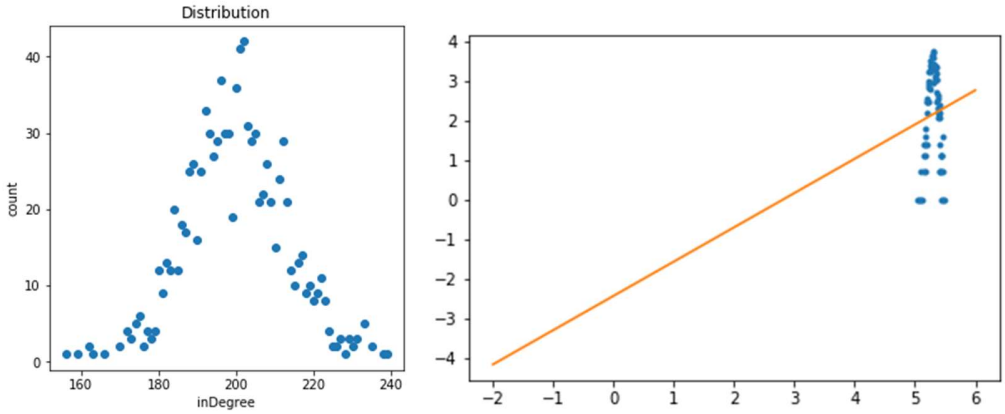
The random graphs with their plots, fitted line to log distribution and slope are as follows:

```
Graph 1(GNM1):Coefficients [-1.09361927  4.69994133] Residuals [9.9345787]
```



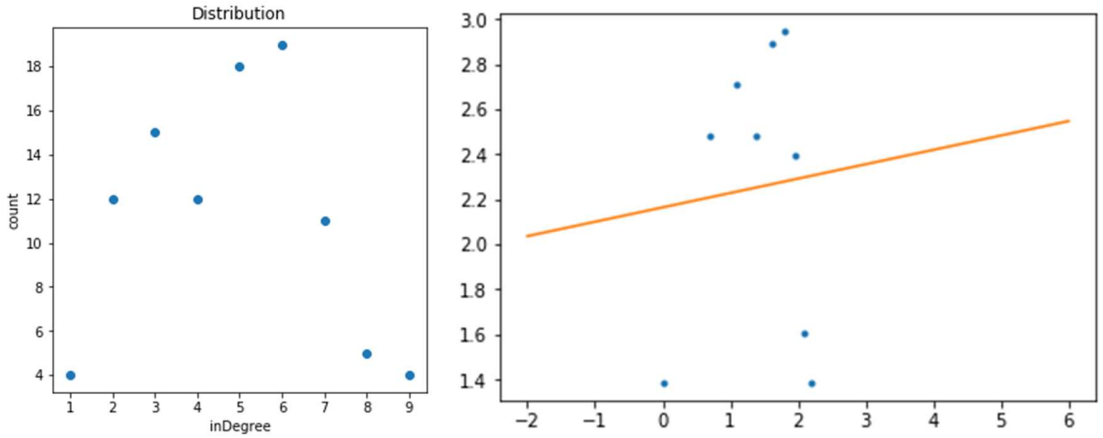
As you can see above, the distribution doesn't follow power law as the log distribution can't be fit with a straight line. Also, the distribution looks like a normal distribution at first glance.

`Graph 2(GNM2):Coefficients[0.86460813 -2.42745382] Residuals [93.11191765]`



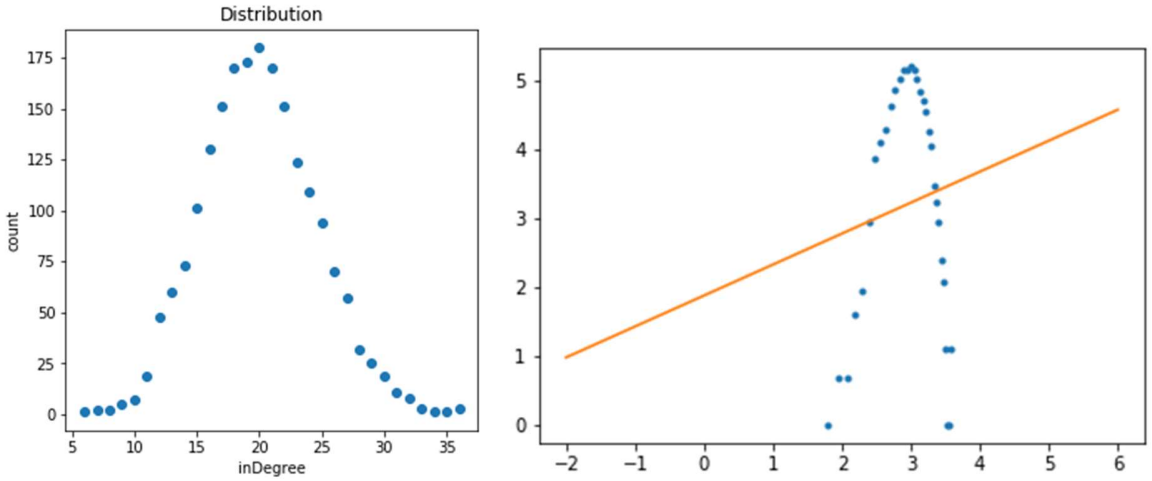
As you can see above, the distribution doesn't follow power law as the log distribution can't be fit with a straight line. Also, the distribution looks like a normal distribution at first glance

`Graph 3(GNP1): Coefficients [0.06379652 2.16398712] Residuals [3.11961529]`



As you can see above, the distribution doesn't follow power law as the log distribution can't be fit with a straight line

`Graph 4(GNP2):Coefficients [0.44868707 1.87506857] Residuals [94.32739034]`



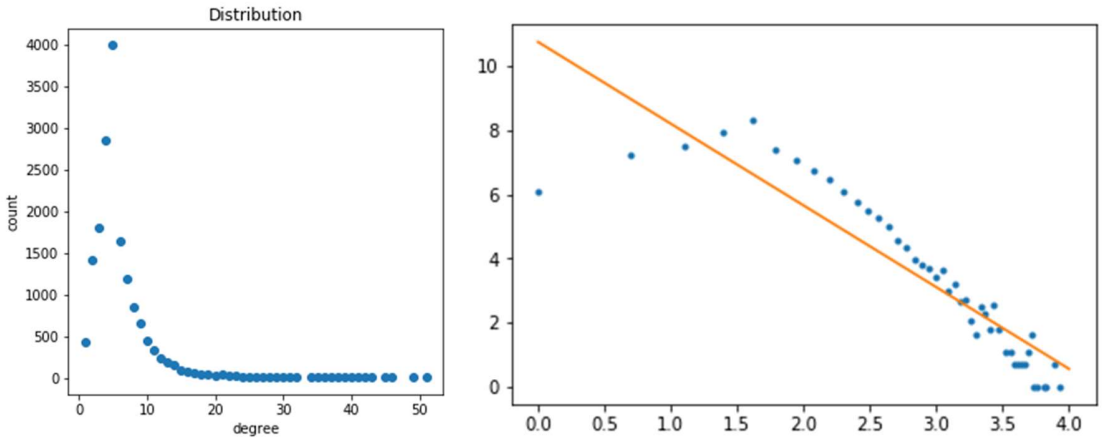
As you can see above, the distribution doesn't follow power law as the log distribution can't be fit with a straight line

Hence, all the randomly generated graphs are not scale-free.

2. Do the Stanford graphs provided to you appear to be scale free?

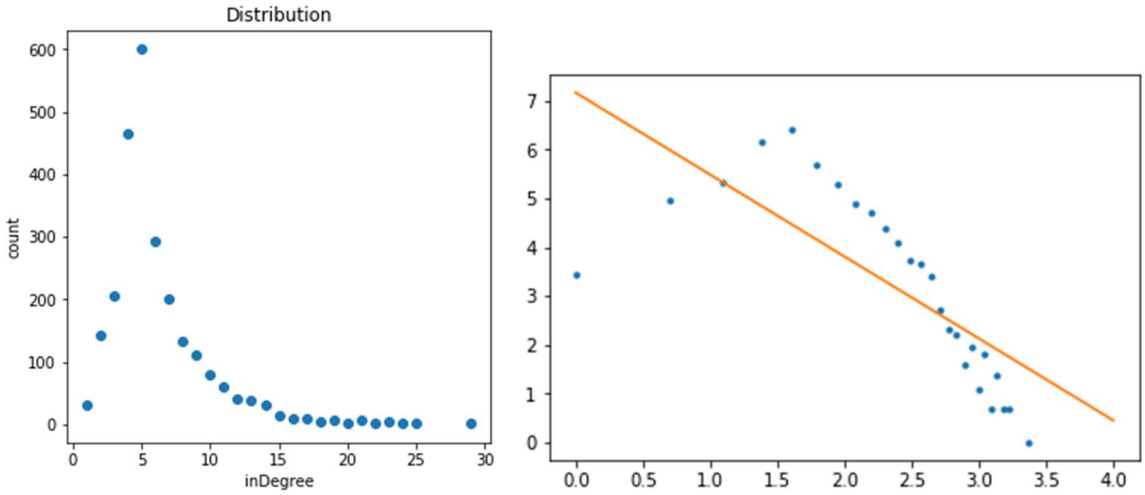
The Stanford graphs provided have the following degree distribution and fit:

1) Amazon Large : Coefficients [-2.54809941 10.75019659] Residuals [53.42809018]



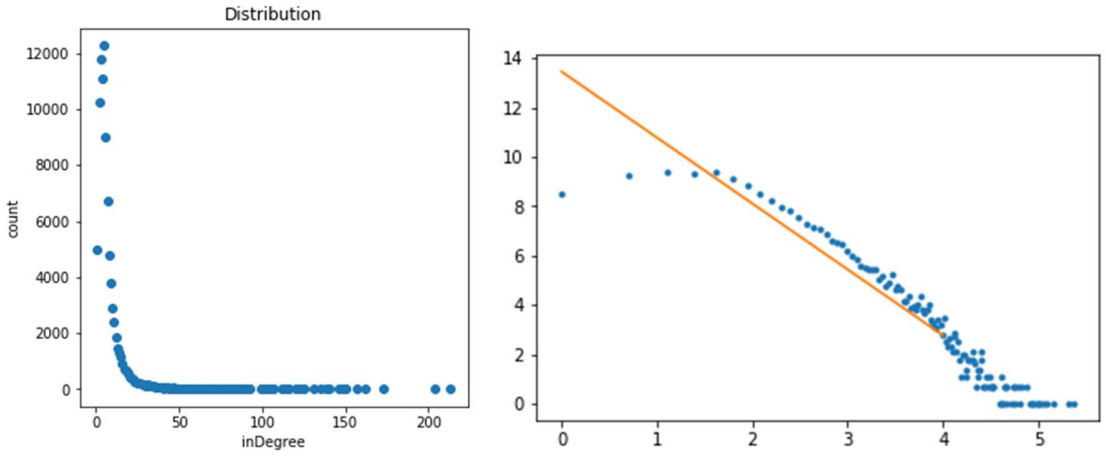
This looks fairly similar to the power distribution except for some values at the start. The fit to the log distribution is also very good. Hence, this graph is scale free.

2) Amazon Small: (Coefficients [-1.67835309 7.16281337] Residuals [39.90749226])



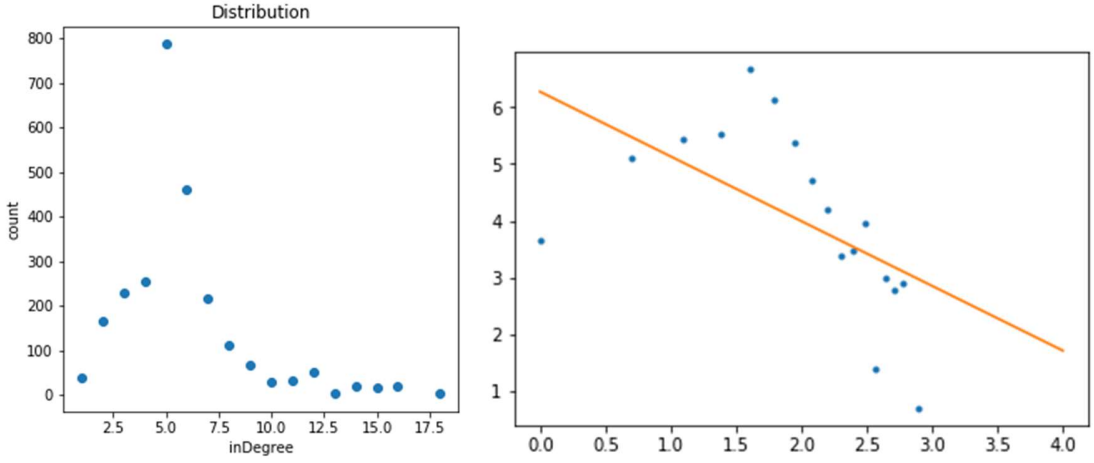
Similar to Amazon Large.

3) DBLP Large: (Coefficients [-2.67185045 13.44832764] Residuals [69.92866355])



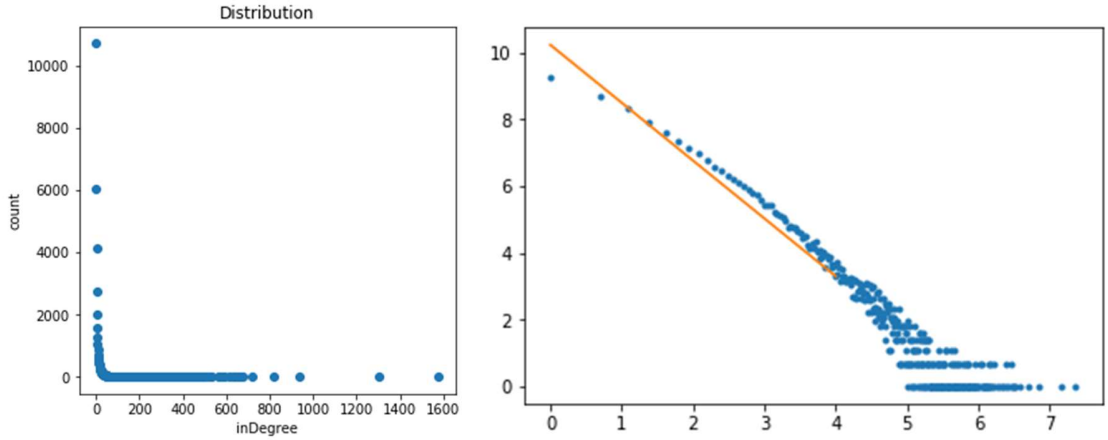
This graph is also very similar to the ideal power distribution and the fit of the straight line in the log distribution is also good. Hence, it is scale free.

4) DBLP Small: (Coefficients [-1.14109288 6.27327106] Residuals [28.71080164])



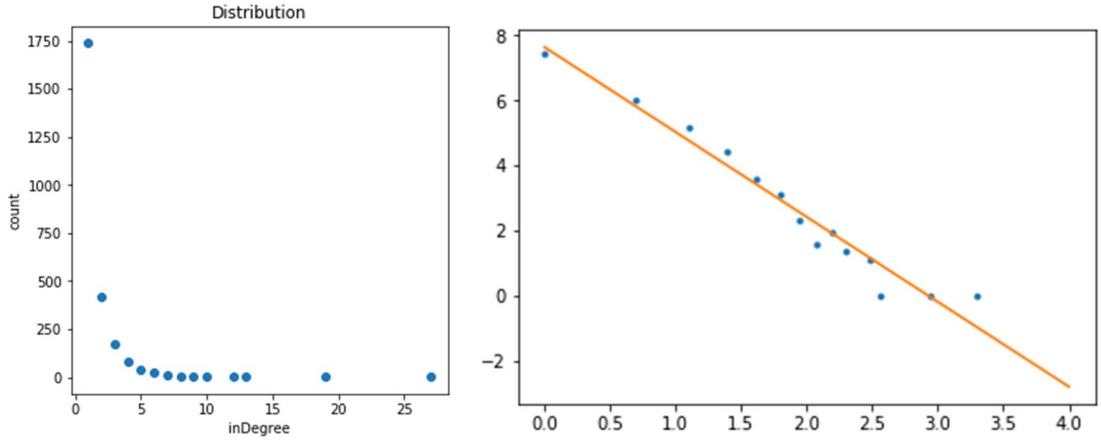
This graph is a little more scattered and has more noise towards the start which causes the fit to skew. However, it still looks like a power distribution to some extent.

5) Youtube Large: (Coefficients [-1.73318997 10.22937638] Residuals [115.5071705])



This graph looks most closest to the ideal power distribution and hence scale free.

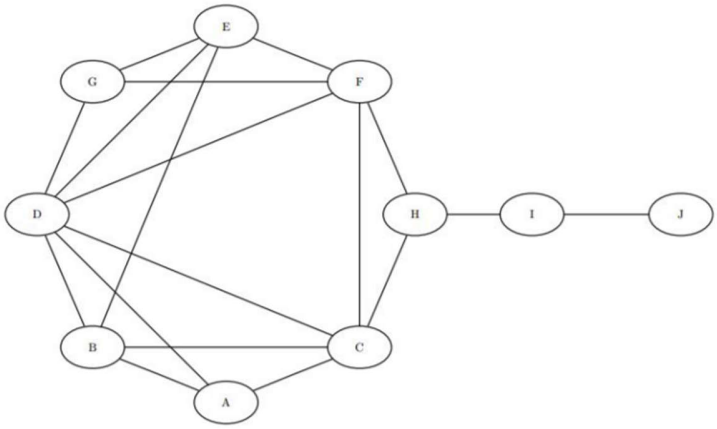
6) Youtube Small: (Coefficients [-2.60472814 7.63227776] Residuals [2.72996778])



Fairly similar to Youtube Large except for less datapoints. Still follows power law.

Hence, the Stanford graphs provided appear to be scale free after looking at the distribution and fit to log distribution.

Problem 2: Centrality



1. Rank the nodes from highest to lowest closeness centrality.

After running our code which finds the shortest path distance of each vertex to all other vertex and then finding the inverse of the sum of all neighbours, the closeness can be found. More the neighbours with lesser distance, more the closeness.

id	closeness
C	0.071428571
F	0.071428571
D	0.066666667
H	0.066666667
B	0.058823529
E	0.058823529
A	0.055555556
G	0.055555556
I	0.047619048
J	0.034482759

2. Suppose we had some centralized data that would sit on one machine but would be shared with all computers on the network. Which two machines would be the best candidates to hold this data based on other machines having few hops to access this data?

The two machines that are connected to most number of computers directly with the least distance will be the best candidates to act as a centralized data store. If we consider the above graph as such a network, then C and F would be the best candidates to store data as they have the highest closeness.

Problem 3: Articulation Points

1. **In this example, which members should have been targeted to best disrupt communication in the organization?**

In the given example of terrorists' network, we can find the people who are most important by finding whether the number of connected components increase by the removal of each person in the network. From the initial connected components of 3, the following terrorists increase the connected components upon their removal and hence the best to target to disrupt their communication.

id	articulation
Usman Bandukra	1
Djamal Beghal	1
Mohamed Atta	1
Mamoun Darkazanli	1
Essid Sami Ben Khemais	1
Raed Hijazi	1
Nawaf Alhazmi	1