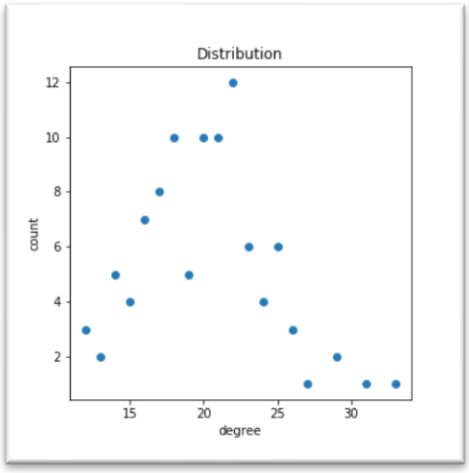
Network Properties in Spark: CSC 591 : ADBI

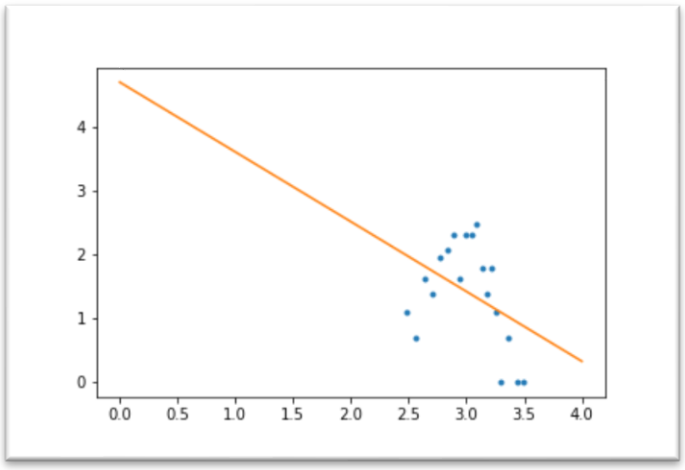
Work Done by: Ashwin Balasubramaniam.

Q1.

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).**

a) GNM1

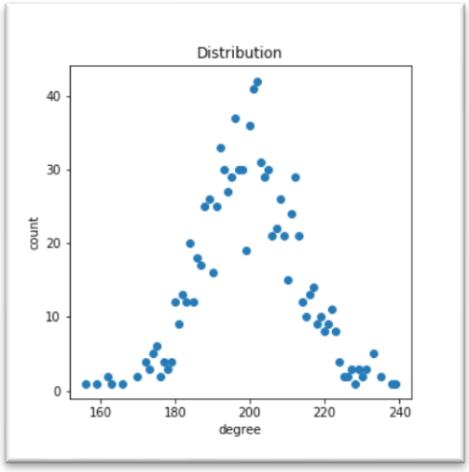


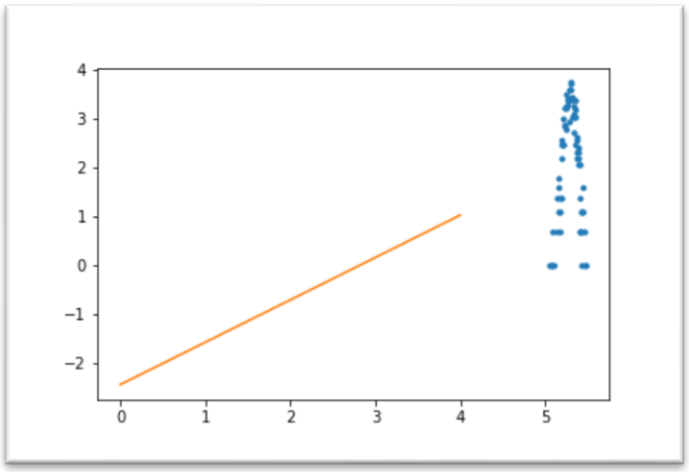


Coefficients [-1.09361927 4.69994133]
Residuals [9.9345787]

From the graph above you can see that the distribution doesn't follow the power law as the log distribution can't be fit with a straight line. Also, the distribution looks normal.

b) GNM2

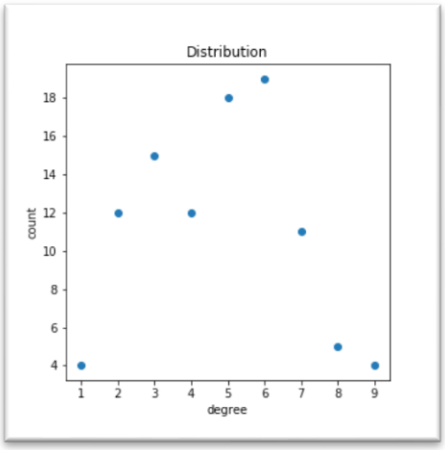


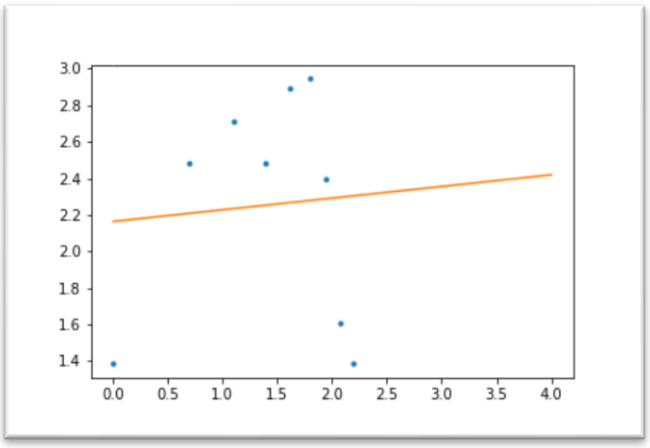


Coefficients [0.86460813 -2.42745382]
Residuals [93.11191765]

From the graph above you can see that the distribution doesn't follow the power law as the log distribution can't be fit with a straight line. Also, the distribution looks normal.

c) GNP1

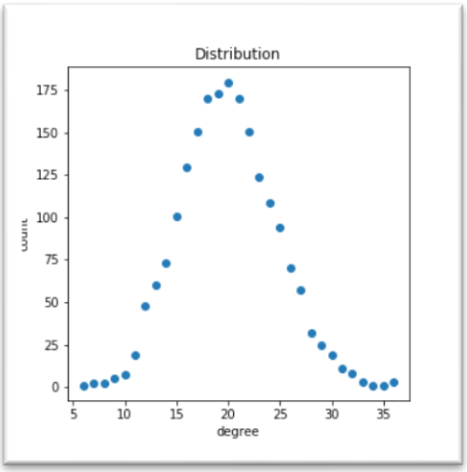


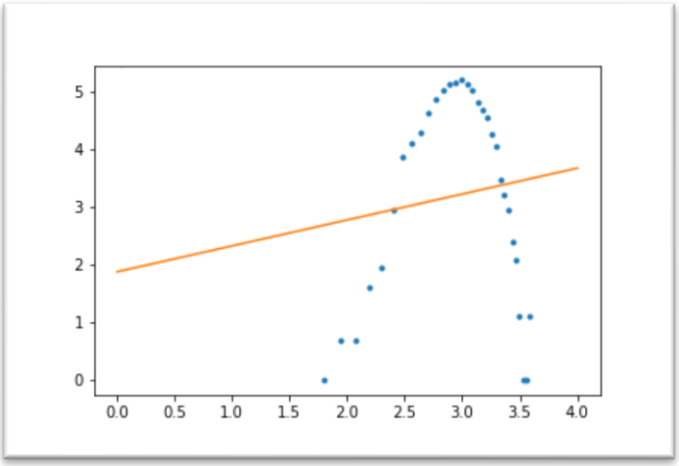


Coefficients [0.06379652 2.16398712]
Residuals [3.11961529]

From the graph above you can see that the distribution doesn't follow the power law as the log distribution can't be fit with a straight line.

d) GNP2





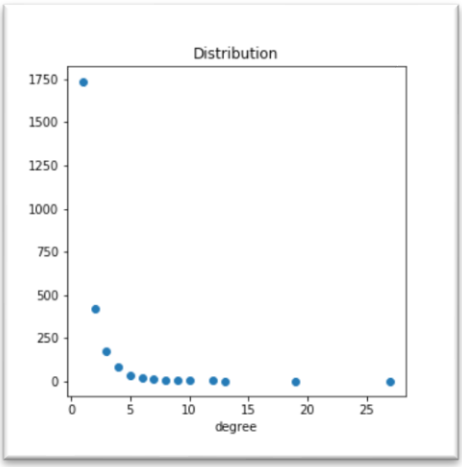
Coefficients [0.44868707 1.87506857]
Residuals [94.32739034]

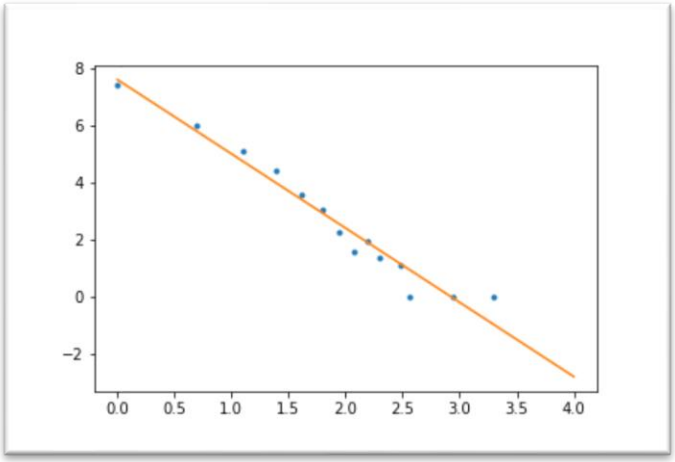
From the graph above you can see that the distribution doesn't follow the power law as the log distribution can't be fit with a straight line.

Hence, it can be concluded that the randomly generated graphs are not scale -free

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

a) Youtube small

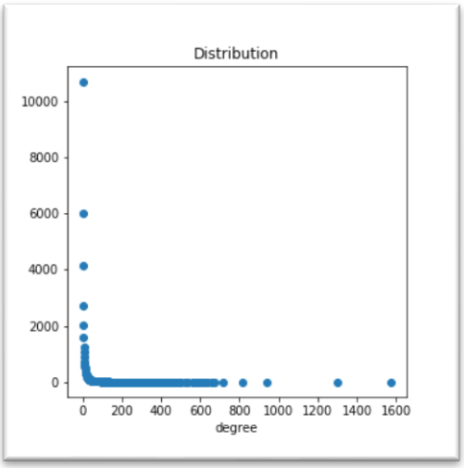


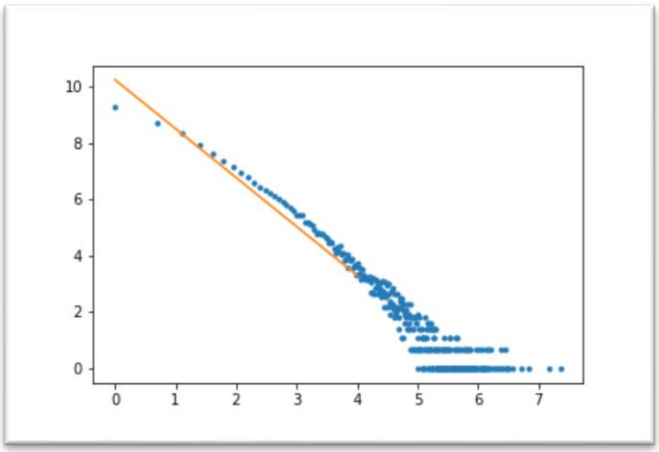


Coefficients [-2.60472814 7.63227776]
Residuals [2.72996778]

From the graphs above it can be seen that the power distribution is almost similar except for the few values. The log distribution also fits good. Hence, the graph can be concluded as scale free

b) Youtube large

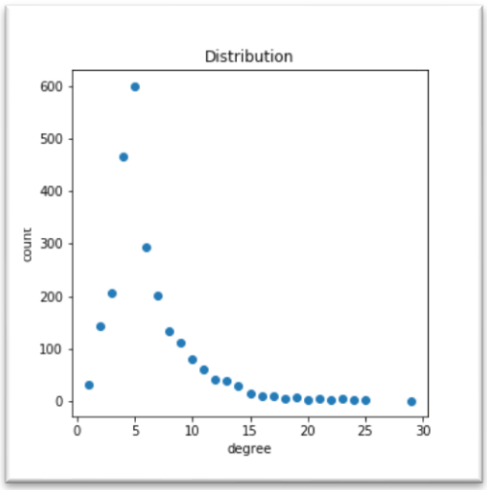


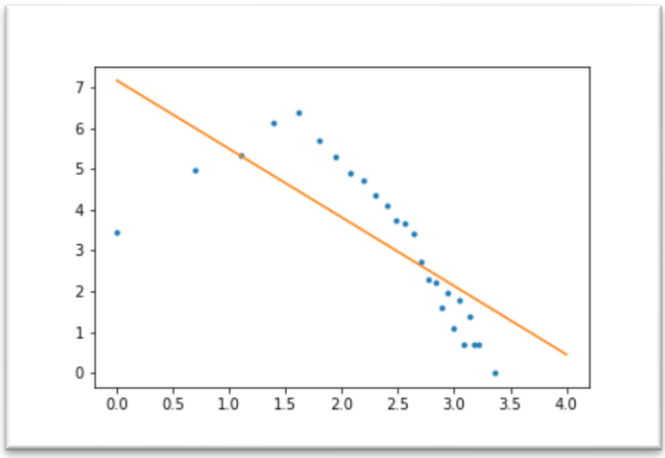


Coefficients [-1.73318997 10.22937638]
Residuals [115.5071705]

The graph appears to be properly distributed in the power distribution and the points like almost on the line. Hence, it can be concluded as scale free

c) Amazon small

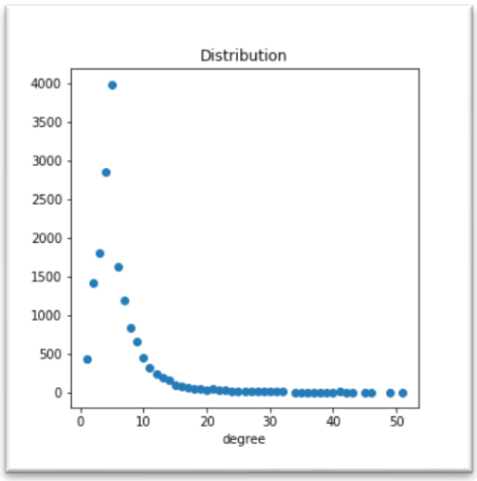


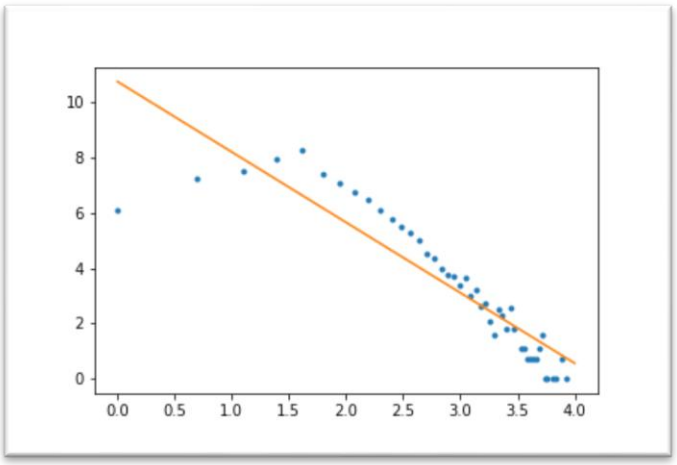


Coefficients [-1.67835309 7.16281337]
Residuals [39.90749226]

From the graphs above it can be seen that the power distribution is almost similar except for the few values. The log distribution also fits good. Hence, the graph can be concluded as scale free

d) Amazon large

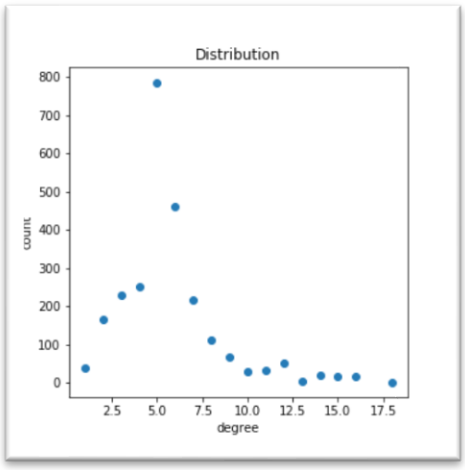


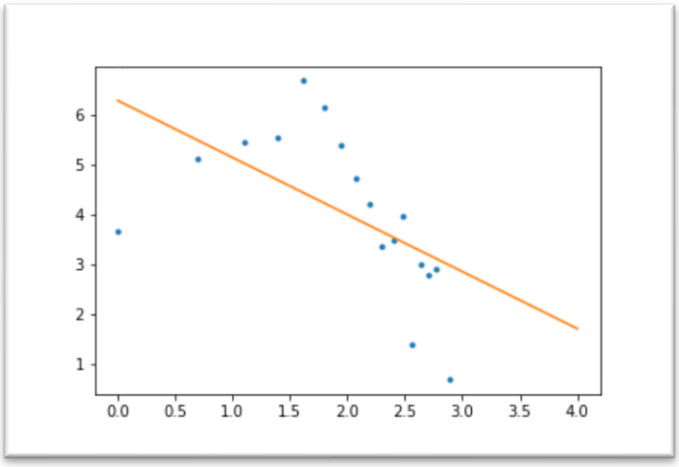


Coefficients [-2.54809941 10.75019659]
Residuals [53.42809018]

From the graphs above it can be seen that the power distribution is almost similar except for the few values. The log distribution also fits good. Hence, the graph can be concluded as scale free

e) DBLP small

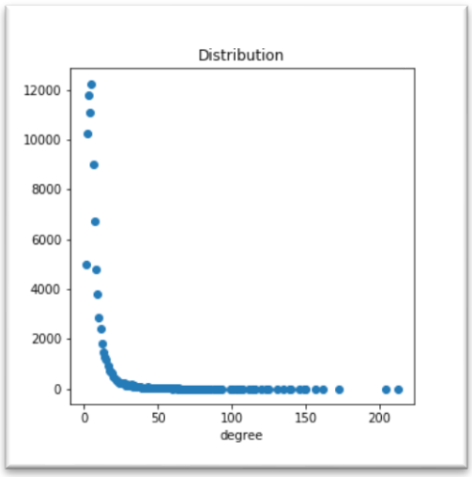


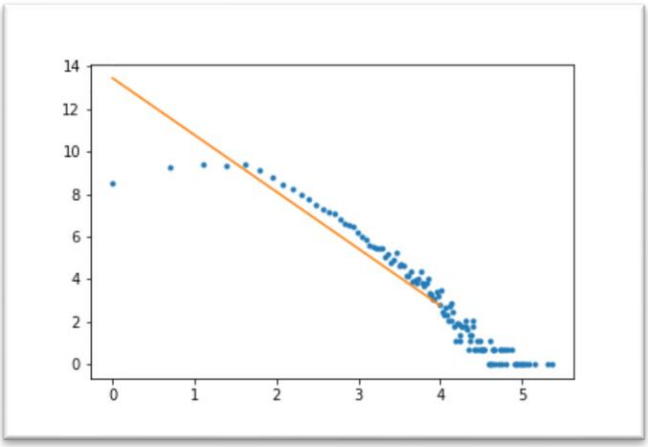


Coefficients [-1.14109288 6.27327106]
Residuals [28.71080164]

The graph appears to be slightly scattered and the scattering is focused mainly in the start which makes the fit to be skewed. Still, it can be concluded that it follow power distribution/

f) DBLP Large





Coefficients [-2.67185045 13.44832764]
Residuals [69.92866355]

From looking at the above graph it can be concluded that the points follow the ideal power distribution and from the second graph it can be seen that the points lie almost on the line. Hence, it can be concluded as scale free.

Hence, the Stanford graphs can be concluded as scale free.

Q2.

1. **Rank the nodes form highest to lowest closeness centrality**.

id	closeness
C	0.071429
F	0.071429
D	0.066667
H	0.066667
B	0.058824
E	0.058824
A	0.055556
G	0.055556
I	0.047619
J	0.034483

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 machines C and F would be the best candidates to hold the database because they have the highest closeness. This means they have the least distance and hence can be used as the centralized data store.

Q3.

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

The members to be removed to disrupt the communications in the organization can be found by removing one member each iteration and checking if the no. of connected components increases. If it increases, then that member is an articulation point and removing him will best disrupt the communication in the organization.

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