Network Properties with Apache Spark

Submitted by: (Krishna Prasanna Bhamidipati) kbhamid2

Script name: powerLaw.py
Instructions to run the script:

1. Make sure the csv file is present to run the graph powerLaw for, is present
2. Type in command prompt: python powerLaw.py <csv file name>
   Eg. python powerLaw.py gnm1.csv
3. The plots can be viewed if the script is run as a python notebook.

Degree Distribution

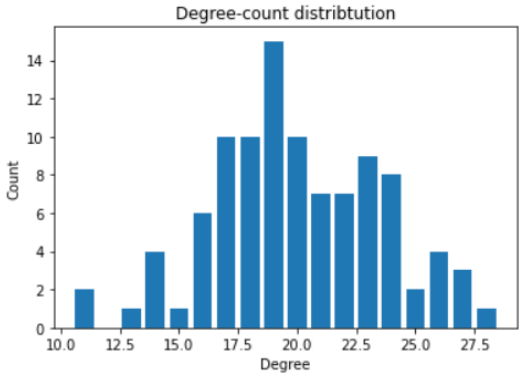
A scale free network is a network that follows power-law distribution. Networks with power-law distributions a called scale-free1 because power laws have the same functional form at all scales.

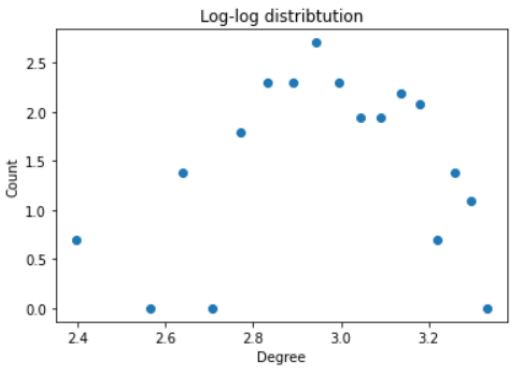
1. Do the random graphs you tested appear to be scale free?

Answer:

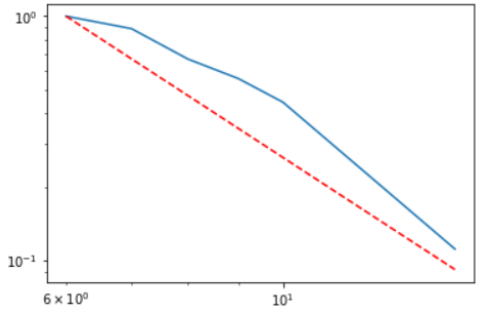
Graph name: gnm1
Alpha value: Alpha: 3.608
Verdict: Does not follow power-law distribution, is not scale-free

```
************************************************
Checking wrt exponential distribution
R value of  -1.3447999302394504 indicates that the distribution is not power-law in nature
p value for this comparison is  0.17868985281353478 which indicates that neither distribution is a significantly strong fit
************************************************
************************************************
Checking wrt log normal distribution
R value of  -0.7775298954447414 indicates that the distribution is not power-law in nature
p value for this comparison is  0.43684620360871596 which indicates that neither distribution is a significantly strong fit
************************************************
```





Complementary cumulative distribution function w.r.t. power-law

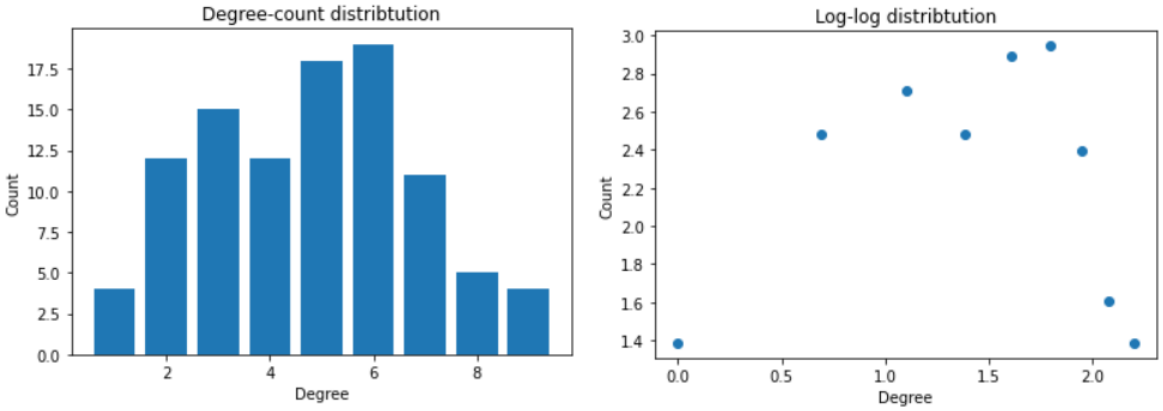


Graph name: gnp1

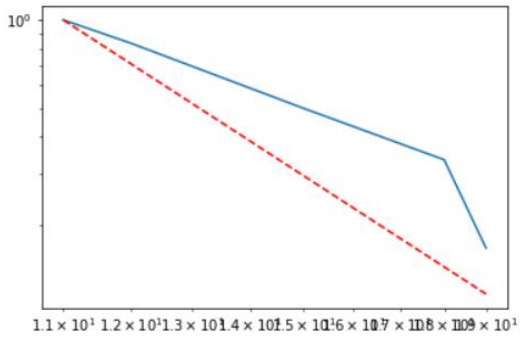
Alpha value: Alpha: 4.939

Verdict: Does not follow power-law distribution, is not scale-free

```
************************************************
Checking wrt exponential distribution
R value of  -0.41533591122277647 indicates that the distribution is not power-law in nature
p value for this comparison is  0.6778960158152955 which indicates that neither distribution is a significantly strong fit
************************************************
************************************************
Checking wrt log normal distribution
R value of  -0.2838481962940842 indicates that the distribution is not power-law in nature
p value for this comparison is  0.7765267175642354 which indicates that neither distribution is a significantly strong fit
************************************************
```



Complementary cumulative distribution function w.r.t. power-law



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

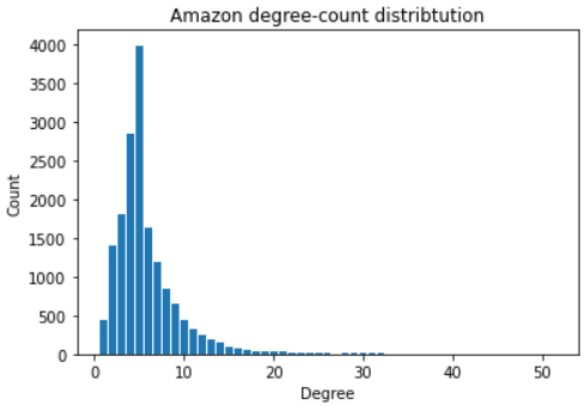
Answer:

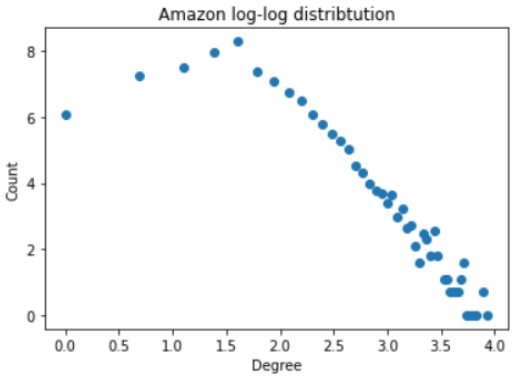
Graph name: Amazon.graph.large

Alpha value: Alpha: 1.325

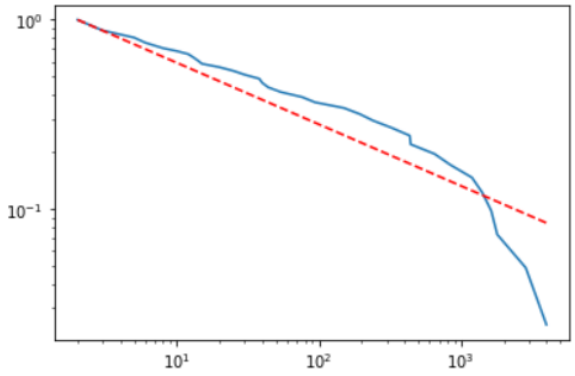
Verdict: Follows power law distribution, is scale free

```
************************************************
Checking wrt exponential distribution
R value of  3.431212684988554 indicates that the distribution is power-law in nature
p value for this comparison is  0.0006008892838471505 which indicates that power-law is a significantly strong fit
************************************************
************************************************
Checking wrt log normal distribution
R value of  -1.1410722354141962 indicates that the distribution is not power-law in nature
p value for this comparison is  0.2538398641384325 which indicates that neither distribution is a significantly strong fit
************************************************
```





Complementary cumulative distribution function w.r.t. power-law

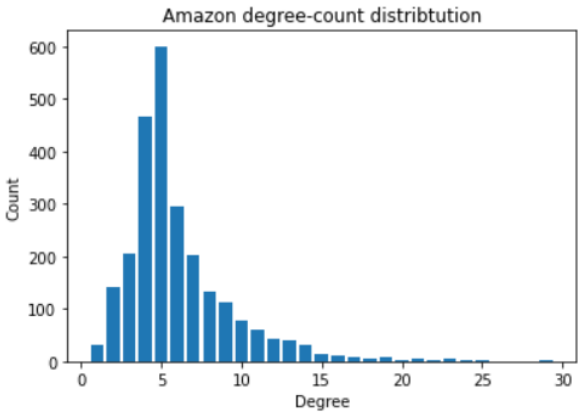


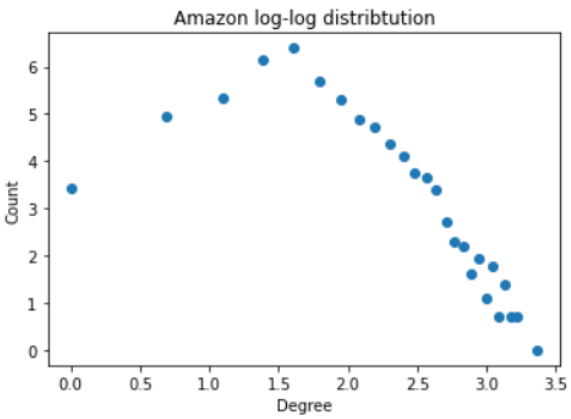
Graph name: Amazon.graph.small

Alpha value: Alpha: 2.3948

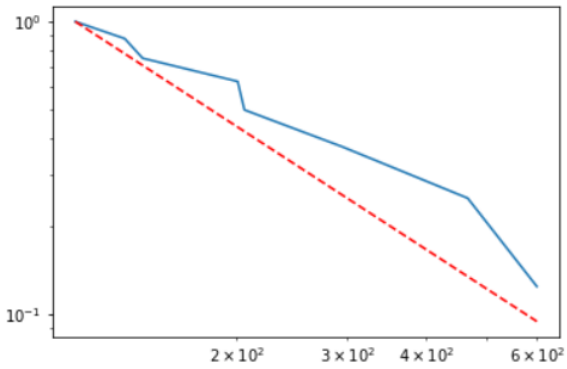
Verdict: Does not follow power law distribution, is not scale free.

```
**************************************************
Checking wrt exponential distribution
R value of  -0.20946280713924925 indicates that the distribution is not power-law in nature
p value for this comparison is  0.8340869669660773 which indicates that neither distribution is a significantly strong fit
**************************************************
**************************************************
Checking wrt log normal distribution
R value of  -0.45407149290414883 indicates that the distribution is not power-law in nature
p value for this comparison is  0.6497773705089622 which indicates that neither distribution is a significantly strong fit
**************************************************
```





Complementary cumulative distribution function w.r.t. powerlaw

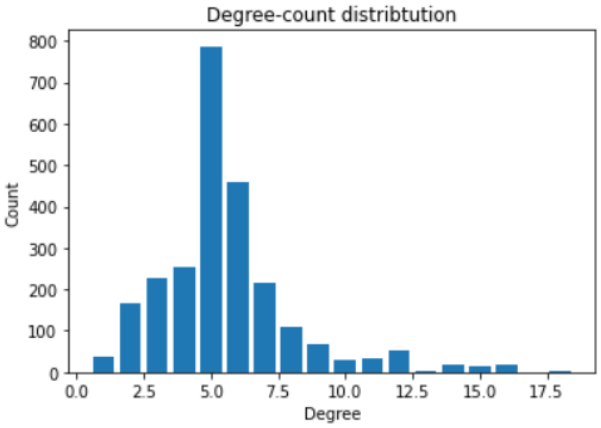


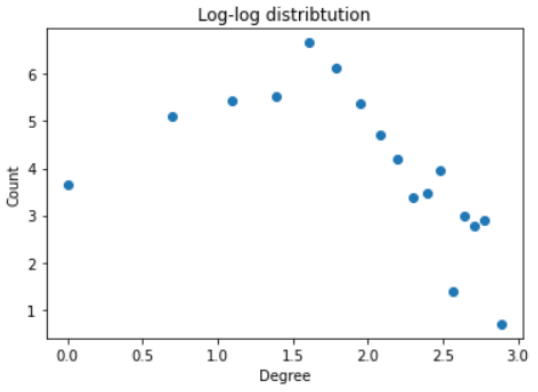
Graph name: dblp.graph.small

Alpha value: Alpha: 1.607

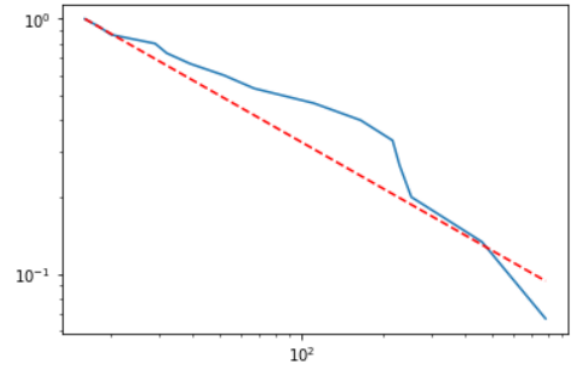
Verdict: Does not follow power law distribution, is not scale free.

```
**************************************************
Checking wrt exponential distribution
R value of  0.37377671372538634 indicates that the distribution is power-law in nature
p value for this comparison is  0.70857044559501 which indicates that neither distribution is a significantly strong fit
**************************************************
**************************************************
Checking wrt log normal distribution
R value of  -0.7793329293156985 indicates that the distribution is not power-law in nature
p value for this comparison is  0.4357836218247241 which indicates that neither distribution is a significantly strong fit
**************************************************
```





Complementary cumulative distribution function w.r.t. powerlaw

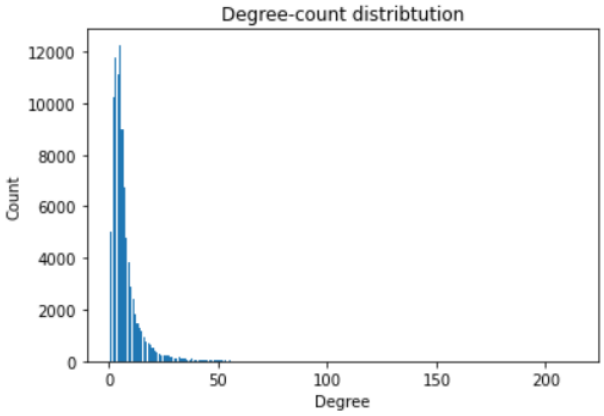


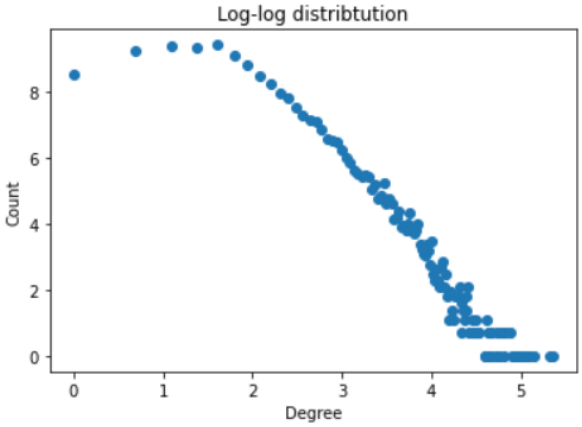
Graph name: dblp.graph.large

Alpha value: Alpha: 1.314

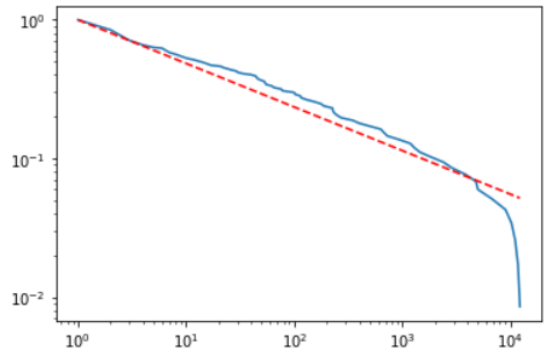
Verdict: Does not follow power law distribution, is not scale free.

```
************************************************
Checking wrt exponential distribution
R value of  9.25779920237234 indicates that the distribution is power-law in nature
p value for this comparison is  2.086891555652333e-20 which indicates that power-law is a significantly strong fit
************************************************
************************************************
Checking wrt log normal distribution
R value of  -1.0657162104438964 indicates that the distribution is not power-law in nature
p value for this comparison is  0.28655194749911195 which indicates that neither distribution is a significantly strong fit
************************************************
```





Complementary cumulative distribution function w.r.t. powerlaw



Centrality:

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

Answer:

id	closeness
F	0.07142857142857142
C	0.07142857142857142
H	0.06666666666666667
D	0.06666666666666667
B	0.058823529411764705
E	0.058823529411764705
A	0.05555555555555555
G	0.05555555555555555
I	0.047619047619047616
J	0.034482758620689655

C, F, H, D, B, E, A, G, I, J

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?

Answer:

The two machines which have highest closeness (i.e. shortest sum of distances) with other machines would be the best candidates to hold such data. In this example, **nodes C and F are the best candidates.**

Articulation Points:

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

Answer:

The members who should have been targeted are – Usman Bandukra, Raed Hijazi, Nawaf Alhazmi, Djamal Beghal, Essid Sami Ben Khemais, Mohamed Atta, Mamoun Darkazanli

Articulation points:

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