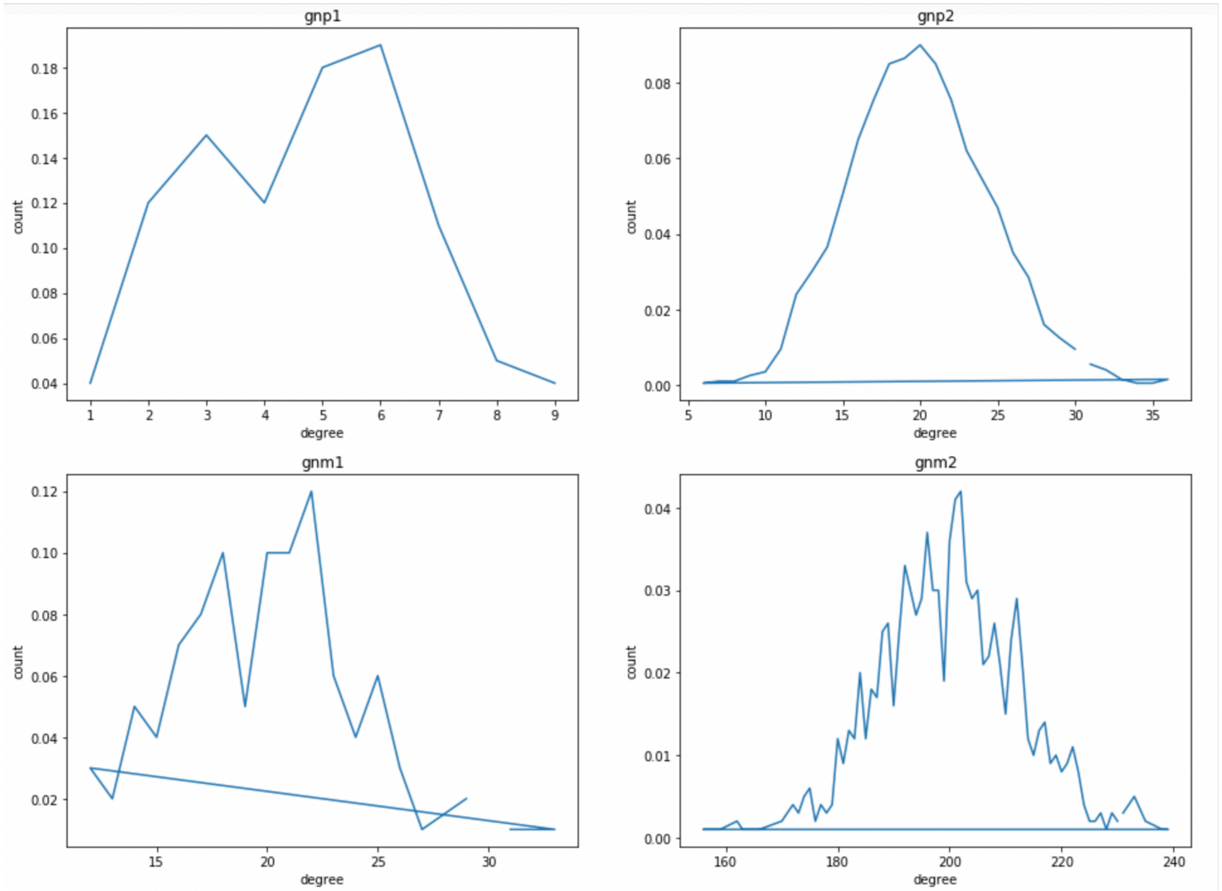
Network Properties In Spark GraphFrames

1. **Do the random graphs you tested appear to be scale free?**
   Following are the random graphs generated.



Gnp1: The power value of graph is: γ = 4.939, which is not in the range of 2 < γ < 3, so it is not scale free.

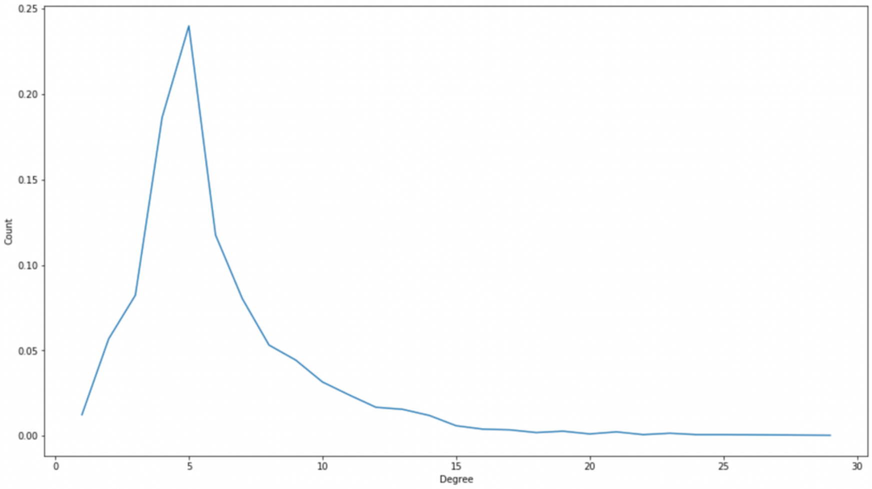
Gnp2: The power value of graph is: γ = 54.5882, which is not in the range of 2 < γ < 3, so it is not scale free.

Gnm1: The power value of graph is: γ = 2.8875, which is in the range of 2 < γ < 3, so it is scale free.

Gnm2: The power value of graph is: γ = 9.620, which is not in the range of 2 < γ < 3, so it is not scale free.
(The γ values are calculated using powerlaw package).

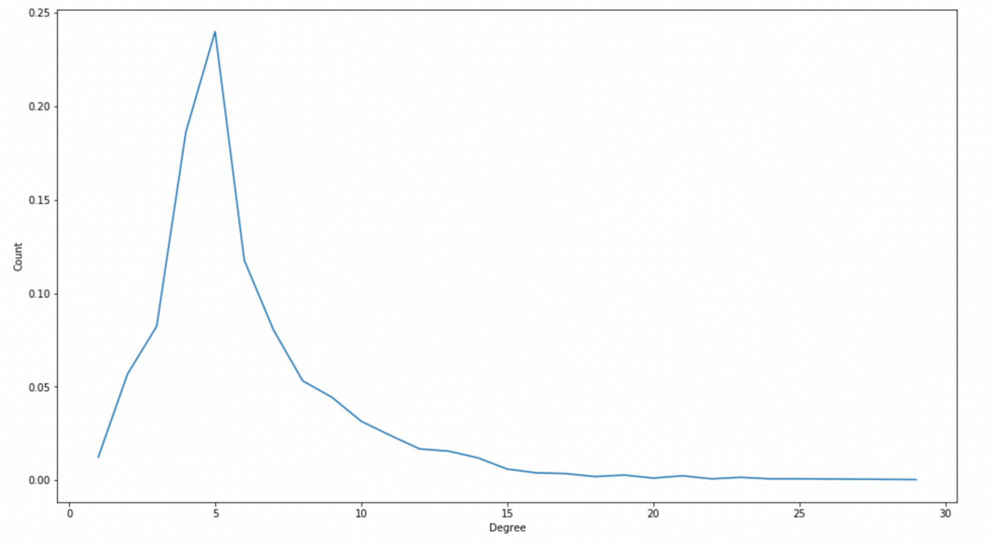
2. **Do the Stanford graphs provided to you appear to be scale free ? Following are the stanford graphs generated.**

Amazon.graph.small:



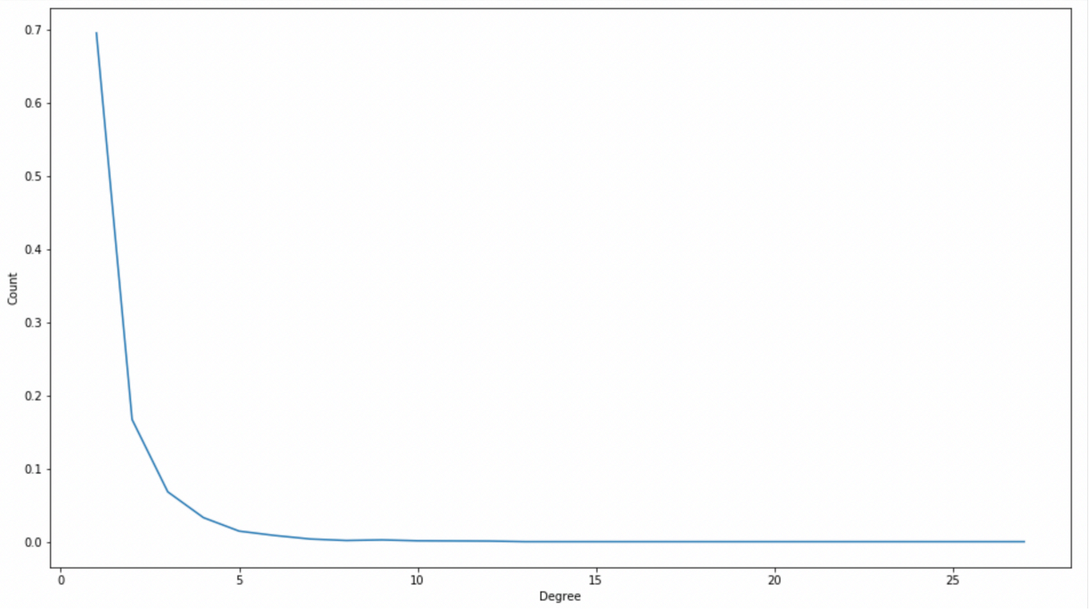
The power value of this graph is: $\gamma = 2.3948$, which is in the range of $2 < \gamma < 3$, so it is scale free.

Amazon.graph.large:



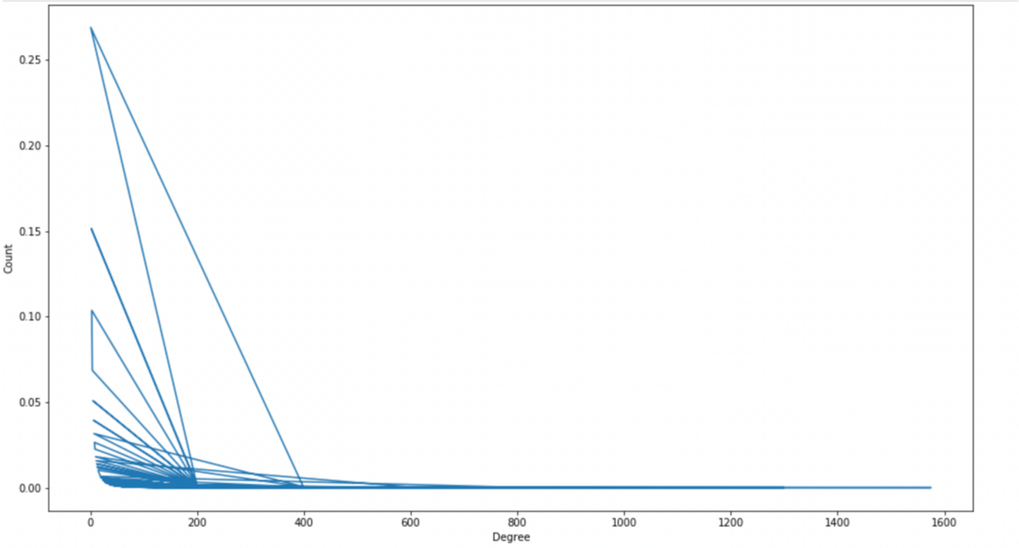
The power value of this graph is: $\gamma = 1.3255$, which is not in the range of $2 < \gamma < 3$, so it is not scale free.

youtube.graph.small:



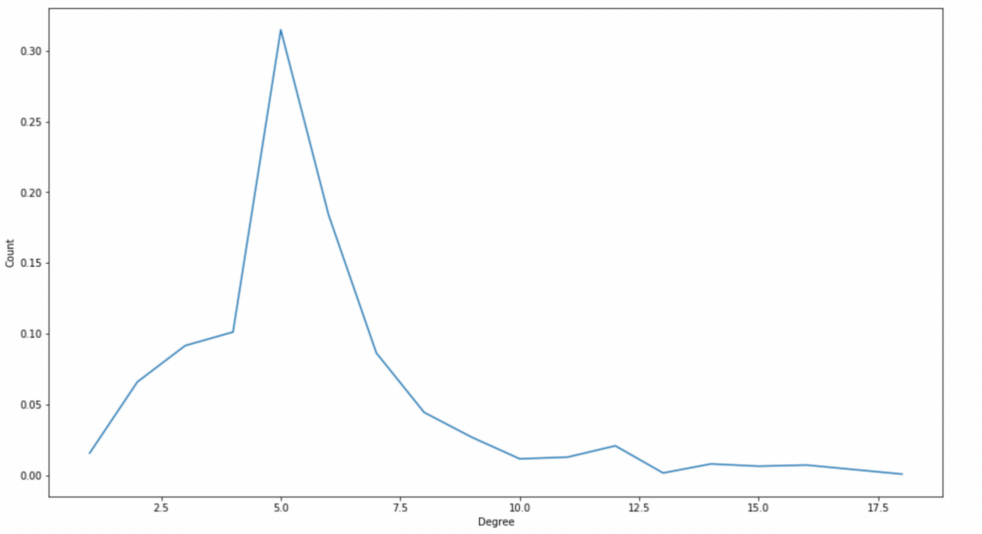
The power value of this graph is: $\gamma = 1.3674$, which is not in the range of $2 < \gamma < 3$, so it is not scale free.

youtube.graph.large:



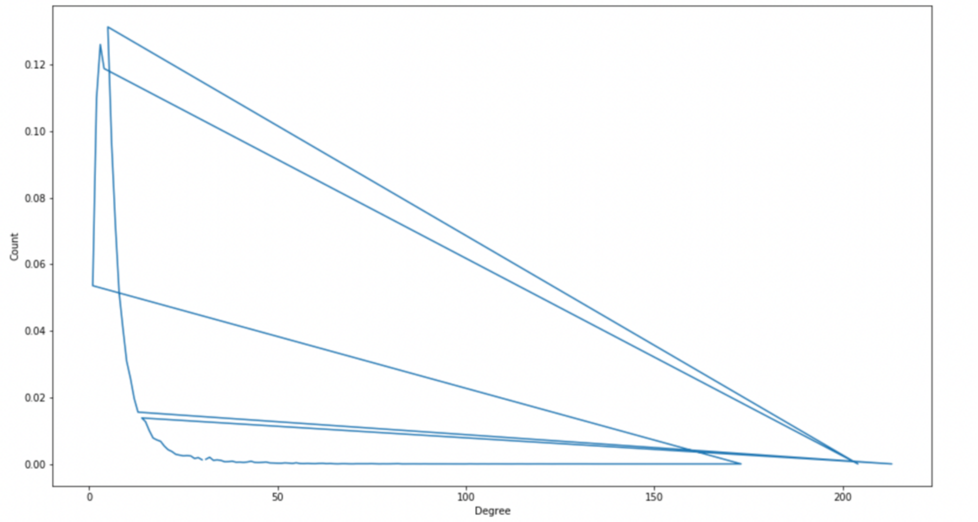
The power value of this graph is: $\gamma = 1.5605$, which is not in the range of $2 < \gamma < 3$, so it is not scale free.

Dblp.graph.small:



The power value of this graph is: γ = 1.6077, which is not in the range of 2 < γ < 3, so it is not scale free.

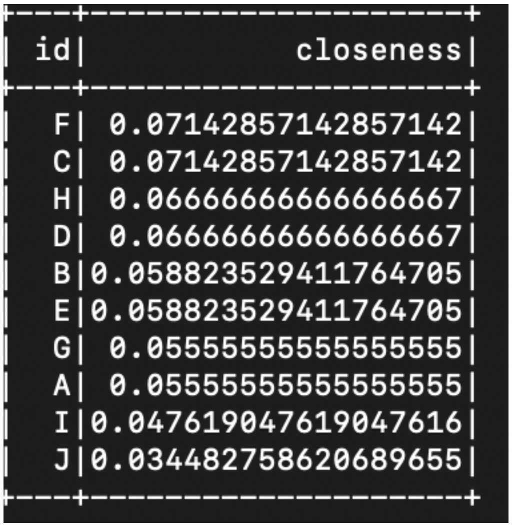
Dblp.graph.large:



The power value of this graph is: γ = 1.3143, which is not in the range of 2 < γ < 3, so it is not scale free.

Centrality

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



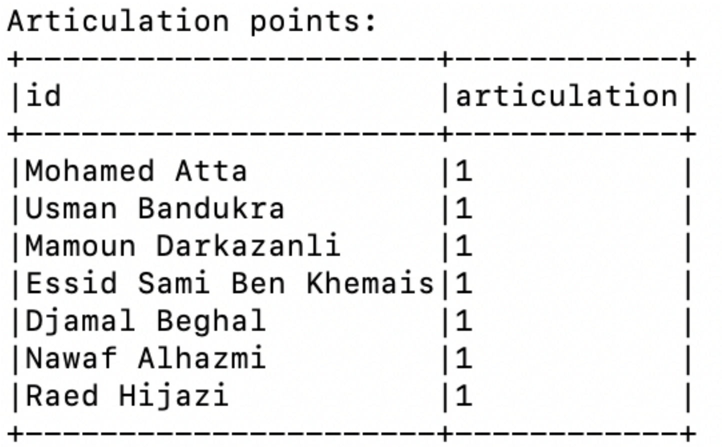
```
+---+--------------------+
| id|           closeness|
+---+--------------------+
|  F| 0.07142857142857142|
|  C| 0.07142857142857142|
|  H| 0.06666666666666667|
|  D| 0.06666666666666667|
|  B|0.058823529411764705|
|  E|0.058823529411764705|
|  G| 0.05555555555555555|
|  A| 0.05555555555555555|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

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 best candidate to hold the data are machines C and F as they have highest closeness value as seen in above table. i.e. The sum of shortest distance from all other nodes to these nodes is least. So other machines will need least number of hops to access these machines.

Articulation

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



```
Articulation points:
+----------------------+------------+
|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           |
+----------------------+------------+
```

To find the members which should be targeted to best disrupt the organization’s communication, we will find out the members which have highest number of connected components in the graph.

The following are the connected components for each member,
Person: Mohamed Atta, Connections: 5
Person: Usman Bandukra, Connections: 4
Person: Mamoun Darkazanli, Connections: 4
Person: Essid Sami Ben Khemais, Connections: 6
Person: Djamal Beghal, Connections: 6
Person: Nawaf Alhazmi, Connections: 4
Person: Raed Hijazi, Connections: 4

Thus, from the above statistics, the members who are connected with highest number of other people are: Essid Sami Ben Khemais and Djamal Beghal. So they should be targeted.