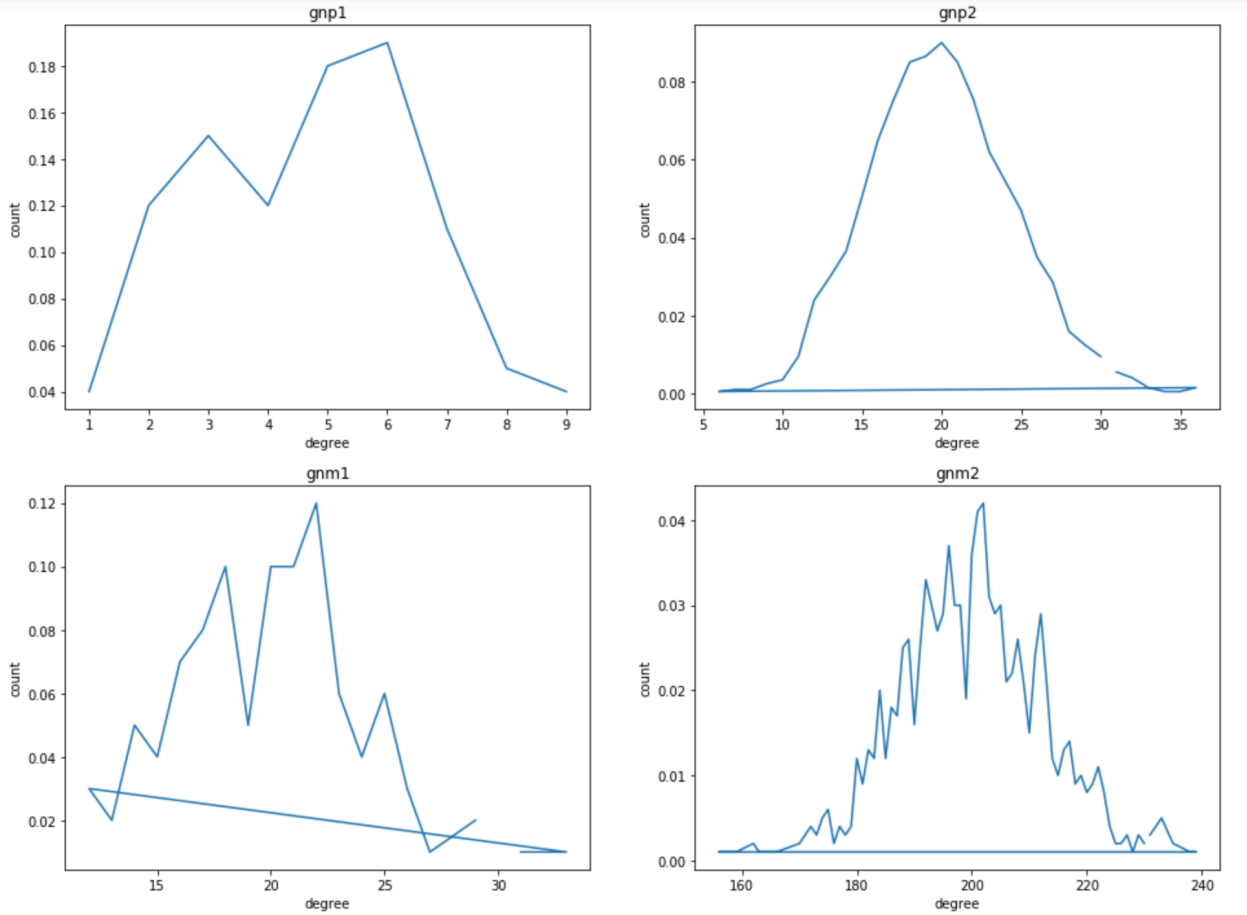
PROJECT: Network Properties In Spark GraphFrames

Generated Random Graphs:



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

Ans: Gnp1: The graph does not appear scale free as its power value is: $\gamma = 4.939$, which is not in the range of $2 < \gamma < 3$.

Gnp2: The graph does not appear scale free as its power value is: γ =54.5882, which is not in the range of $2 < \gamma < 3$.

Gnm1: The graph does appear scale free as its power value is: γ =2.8875, which is in the range of $2 < \gamma < 3$.

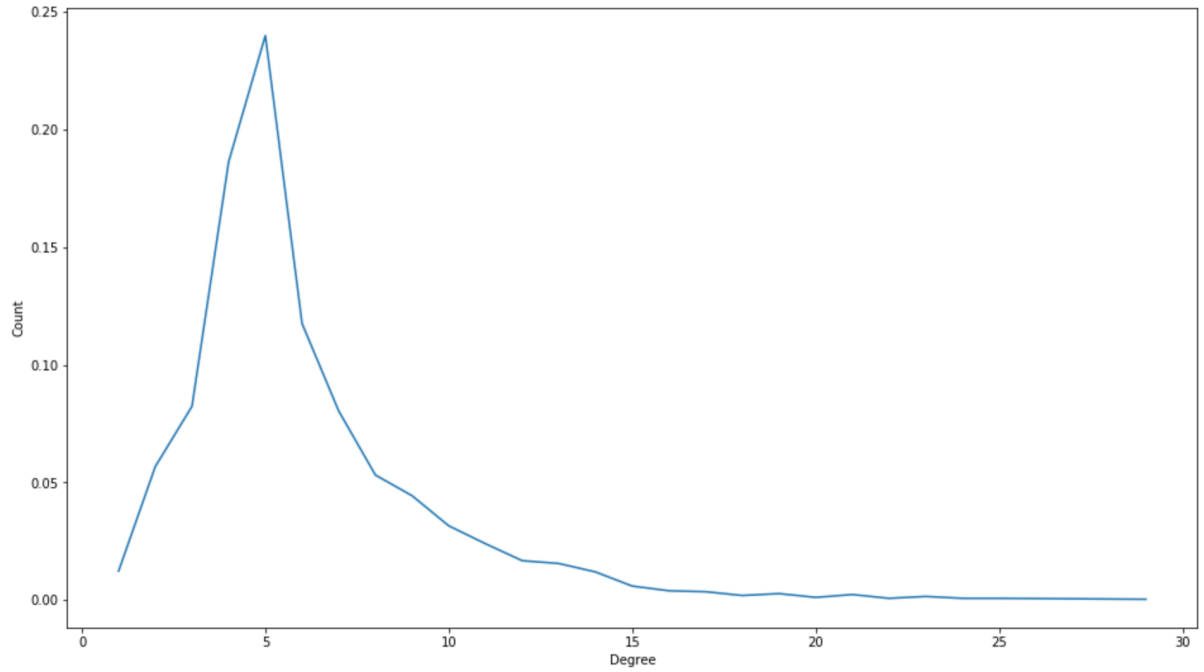
Gnm2: The graph does not appear scale free as its power value is: γ =9.620, which is not in the range of $2 < \gamma < 3$.

(The γ values are calculated using powerlaw package).

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

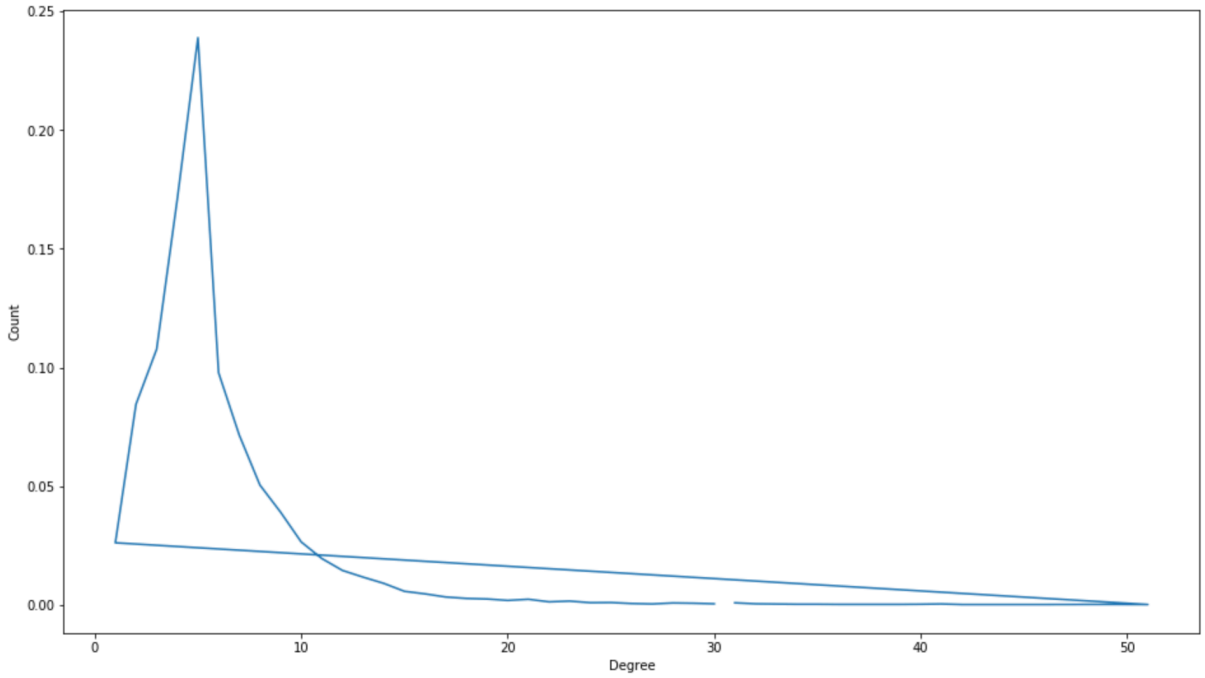
Answer)

Amazon.graph.small:



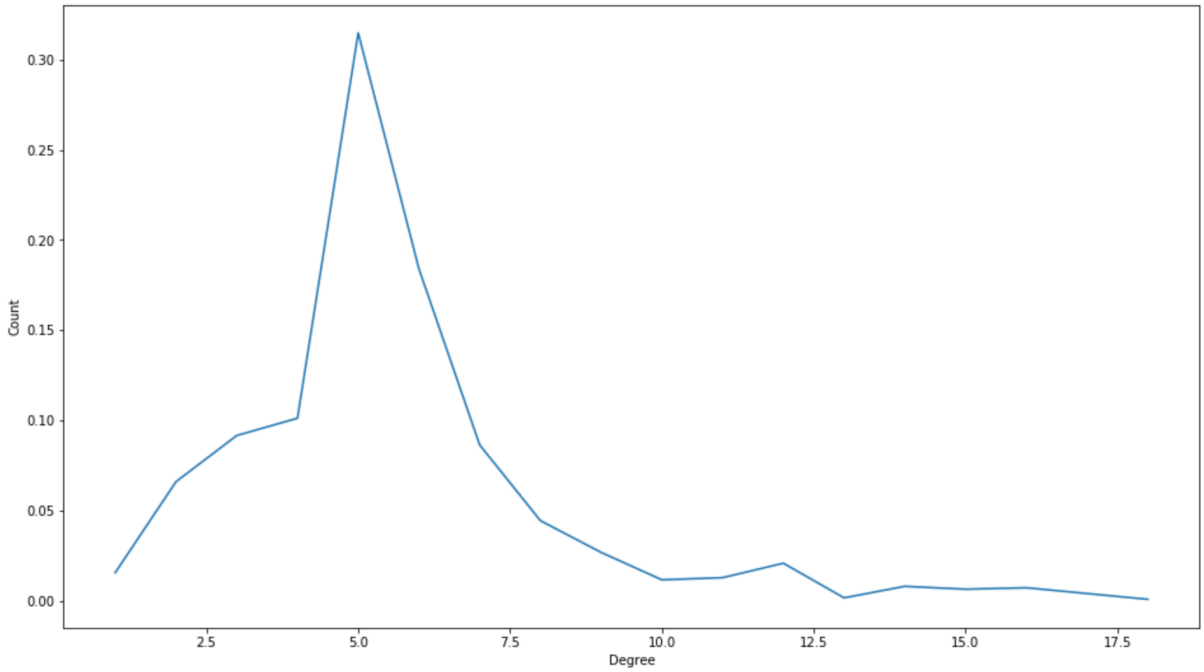
The graph does appear scale free as its power value is: $\gamma = 2.3948$, which is in the range of $2 < \gamma < 3$.

Amazon.graph.large:

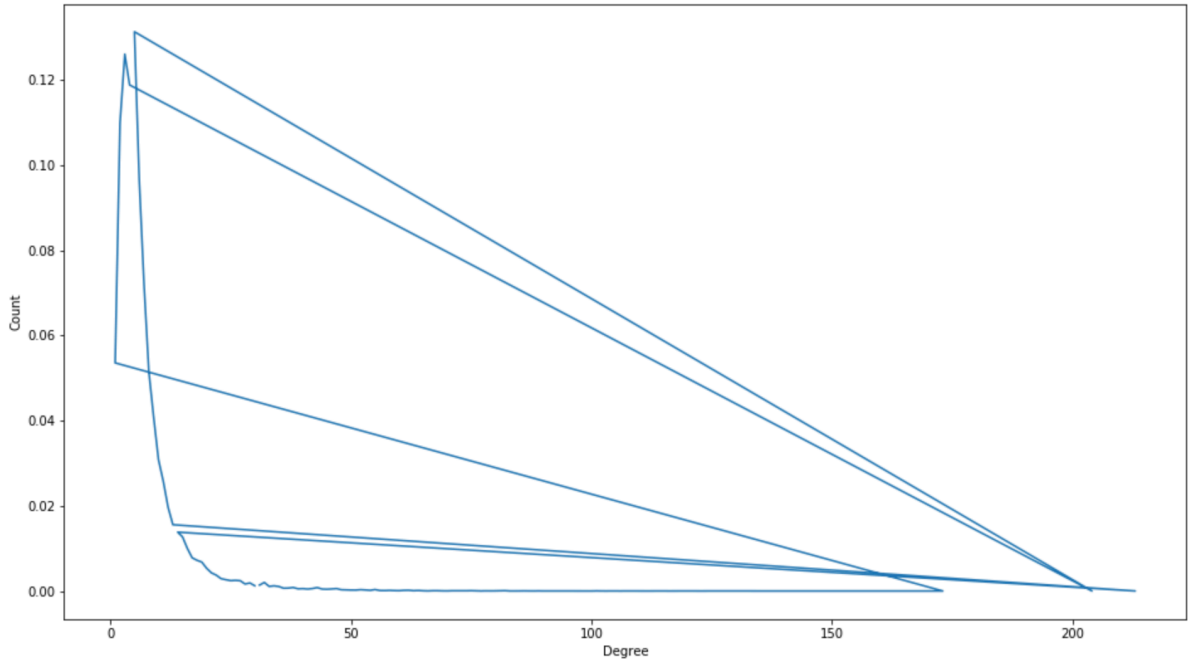


The graph does not appear scale free as its power value is: $\gamma = 1.3255$, which is not in the range of $2 < \gamma < 3$.

Dblp.graph.small:

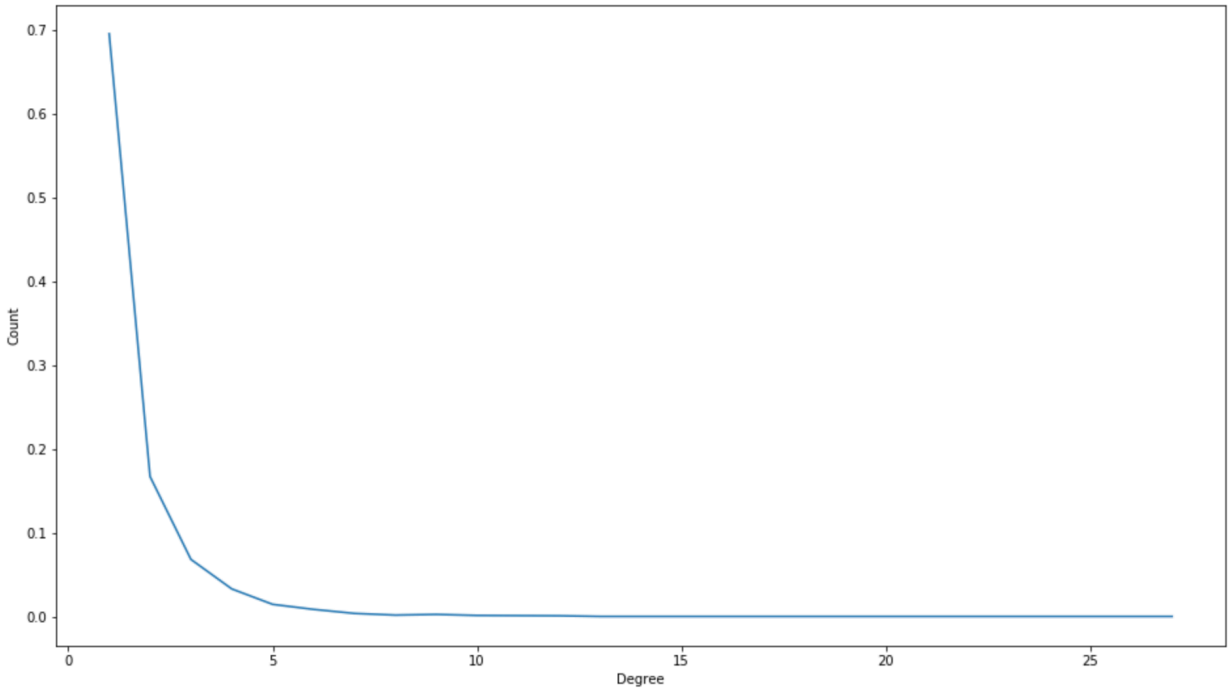


The graph does not appear scale free as its power value is: $\gamma = 1.6077$, which is not in the range of $2 < \gamma < 3$.
Dblp.graph.large:

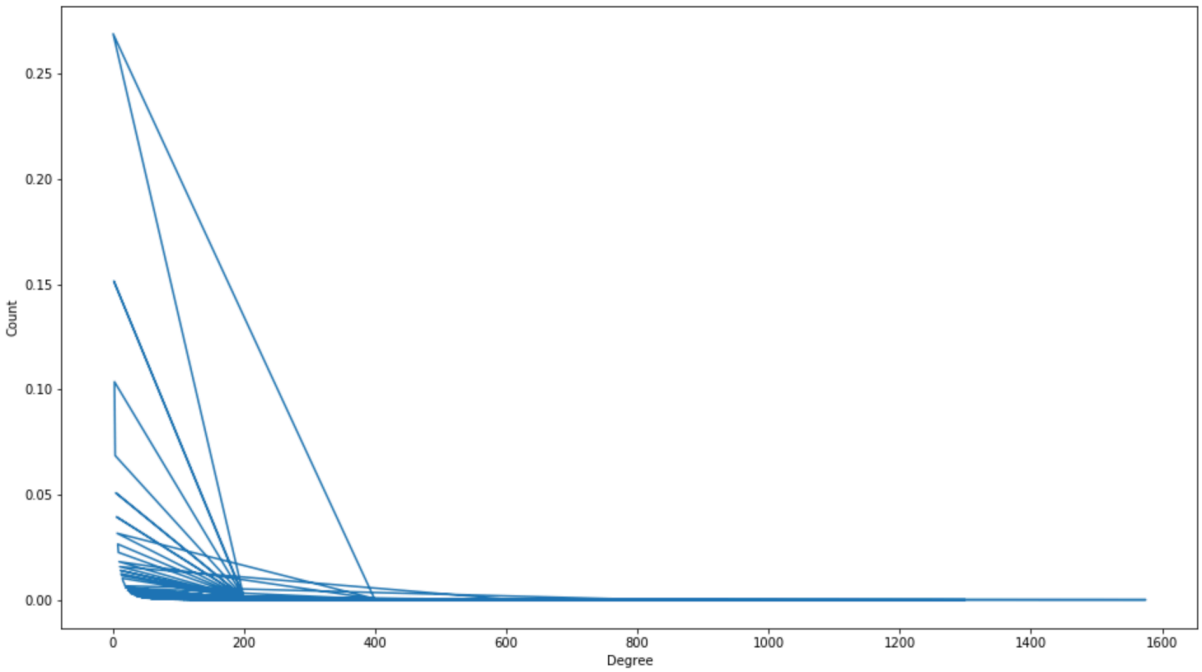


The graph does appear scale free as its power value is: $\gamma = 1.3143$, which is not in the range of $2 < \gamma < 3$ but very near to it.

youtube.graph.small:



The graph does not appear scale free as its power value is: $\gamma = 1.3674$, which is not in the range of $2 < \gamma < 3$.
youtube.graph.large:



The graph does not appear scale free as its power value is: $\gamma = 1.5605$, which is not in the range of $2 < \gamma < 3$.

Centrality

Q.1. Rank the nodes from highest to lowest closeness centrality.
Ans)

```
Calculating closeness:
+---+--------------------+
| 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|
+---+--------------------+
```

Q.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 machines C and F would be the best candidate to hold the data as they have the highest closeness value. This means that the sum of shortest from all other nodes to these nodes is the least. Hence, other machines need least number of hops to access these machines

Articulation:

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

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

In order to find members which should be targeted to best disrupt the organization's communication we'll find out the members which has the highest number of connected components in the graph, i.e. the person who is connected with highest number of other people:

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, the person who is connected with highest number of other people are: Essid Sami Ben Khemais and Djamal Beghal