Network Properties in Spark GraphFrames

Degree Distribution

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).
Solution:

1. gnm1
   By using the powerlaw function:
   $\gamma = 2.8875$
   It lies in the range $2 < \gamma < 3$. Hence this graph appears to be scale free.
2. gnm2
   By using the powerlaw function:
   $\gamma = 9.6206$
   It lies in the range $2 < 3 < \gamma$. Hence this graph is not scale free.
3. gnp2
   By using the powerlaw function:
   $\gamma = 54.58$
   It lies in the range $2 < 3 < \gamma$. Hence this graph is not scale free.
4. gnp1
   By using the powerlaw function:
   $\gamma = 4.9390$
   It lies in the range $2 < 3 < \gamma$. Hence this graph is not scale free.

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

1. amazon.graph.small
   By using the powerlaw function:
   $\gamma = 2.3948$
   It lies in the range $2 < \gamma < 3$. Hence this graph appears to be scale free.
2. amazon.graph.large
   $\gamma = 1.3255$
   It lies in the range $\gamma < 2 < 3$. Hence this graph is not scale free.
3. dblp.graph.small
   $\gamma = 1.6077$
   $\gamma < 2 < 3$. Therefore, this graph is not scale free.

4. dblp.graph.large

   $\gamma = 1.3143$

   It lies in the range $\gamma < 2 < 3$. Hence this graph is not scale free.
5. youtube.graph.small

   $\gamma = 1.367$

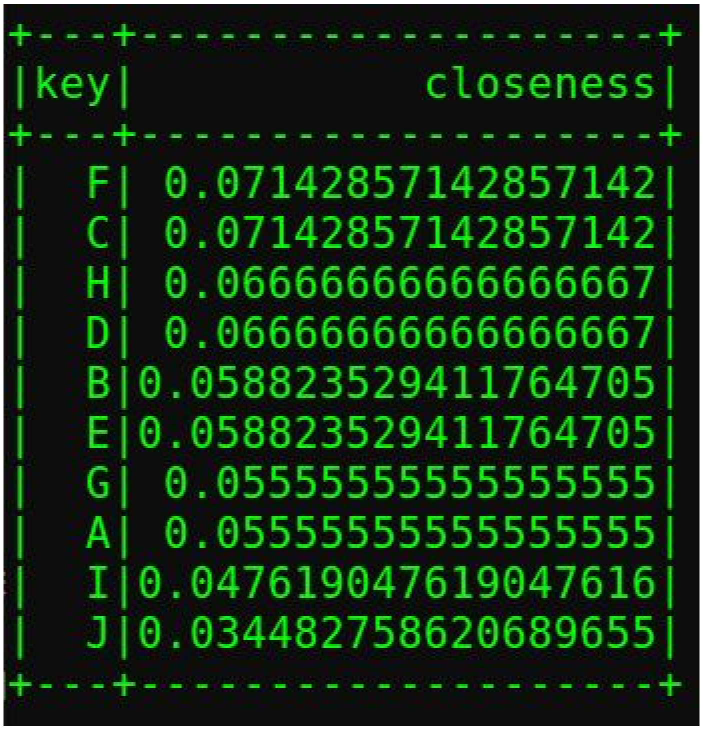
   It lies in the range $\gamma < 2 < 3$. Hence this graph is not scale free.
6. youtube.graph.large

   $\gamma = 1.5605$

   It lies in the range $\gamma < 2 < 3$. Hence this graph is not scale free.

Centrality:

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



```
+---+--------------------+
|key|           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|
+---+--------------------+
```

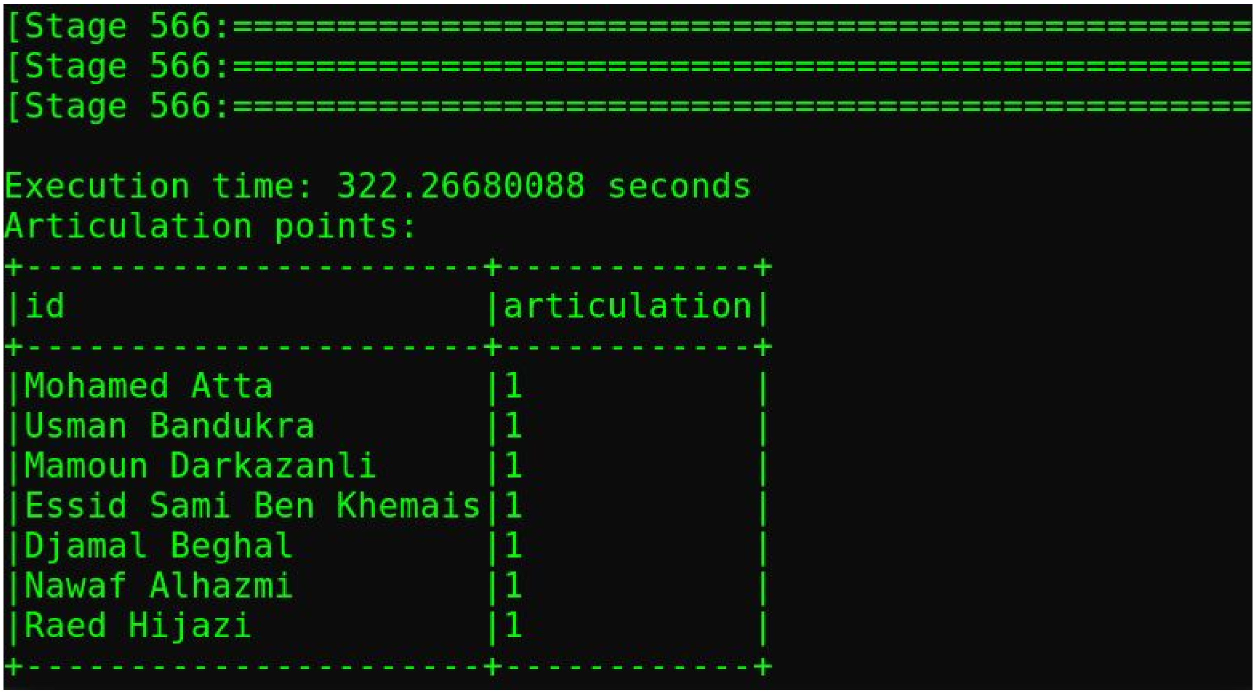
2. Suppose we had some centralized data that would sit on one machine but would be shared with all computers on the network. Which two machines would be the best candidates to hold this data based on other machines having few hops to access this data?

Solution:

The vertex F and C would be a good choice as they have the highest closeness as compared to the other vertices of the graph.

Articulation Points

1.



```
[Stage 566:=================================================
[Stage 566:=================================================
[Stage 566:=================================================

Execution time: 322.26680088 seconds
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           |
+----------------------+------------+
```