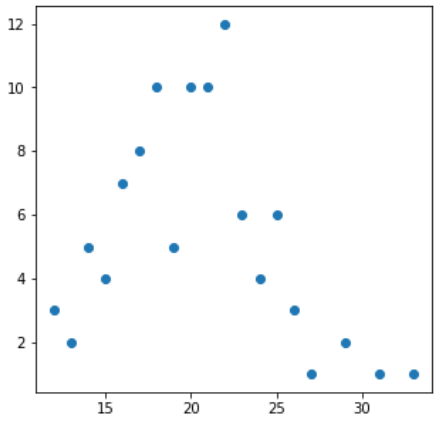
Project Title: Network Properties with Apache Spark
CSC 591 – ADBI

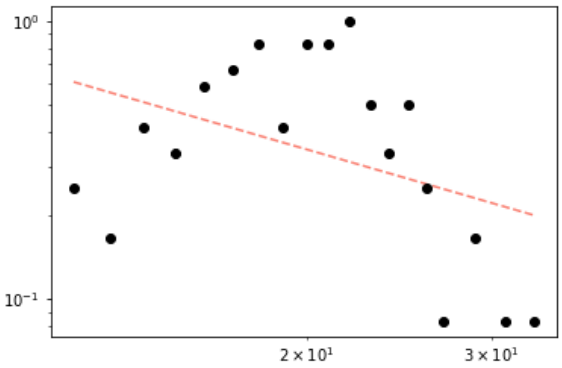
Name: Rohan Pillai
Unity ID: rspillai

Degree Distribution

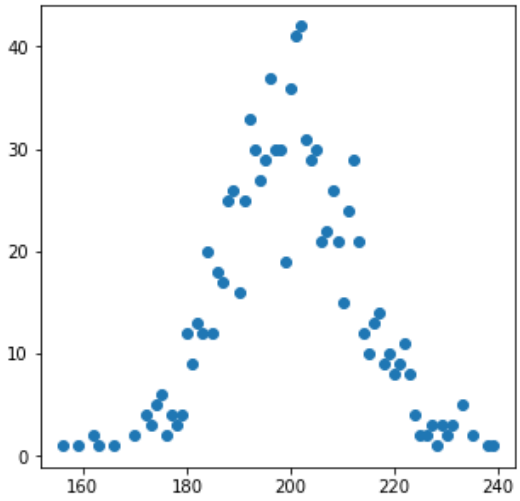
1. Generate a few random graphs. You can do this using networkx's random graph generators. Do the random graphs you tested appear to be scale free? (Include degree distribution with your answer).

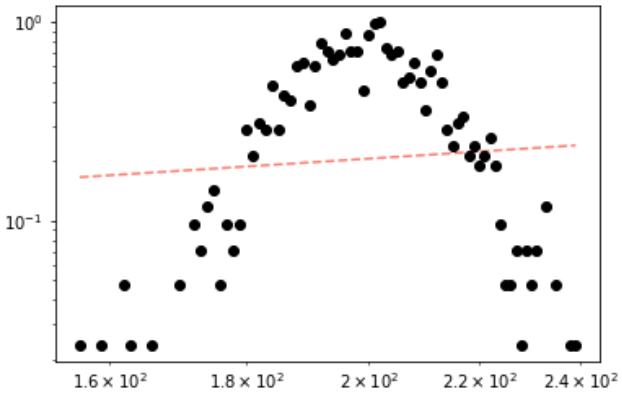
   GNM1
   This graph is not scale free as it does not follow power law.





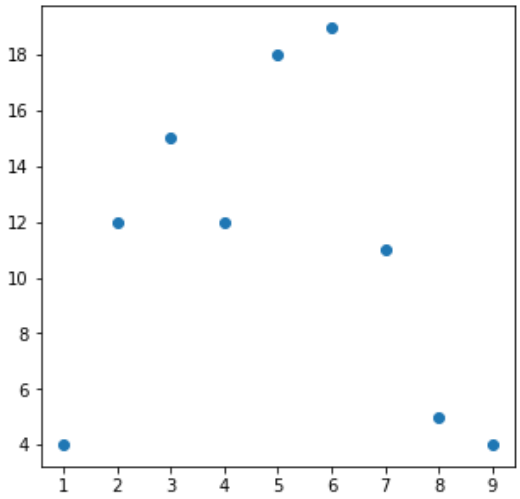
   GNM2
   This graph is not scale free as it does not follow power law.

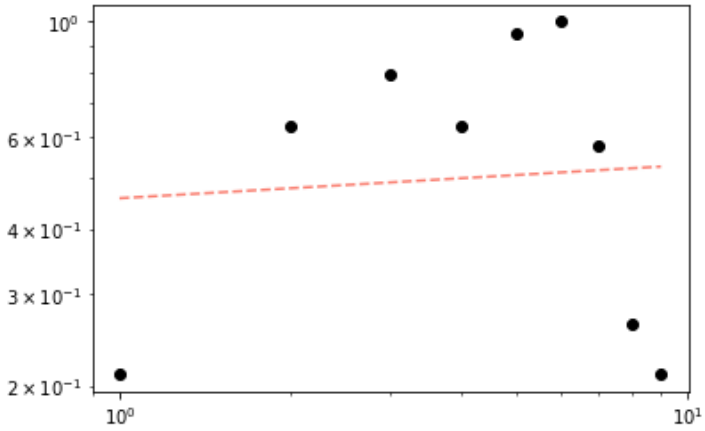




   GNP1

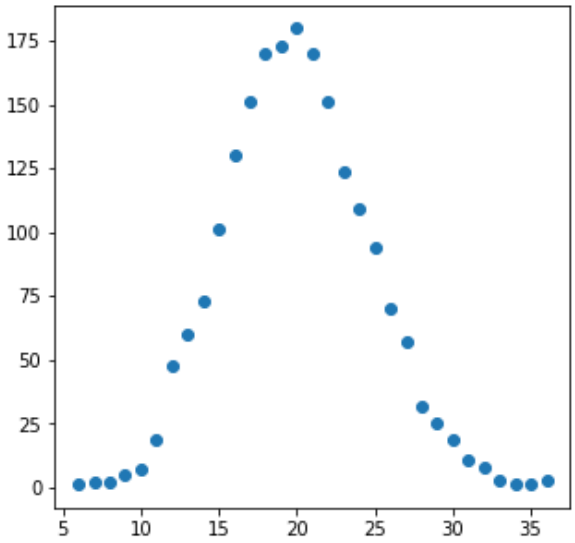
This graph is not scale free as it does not follow power law.

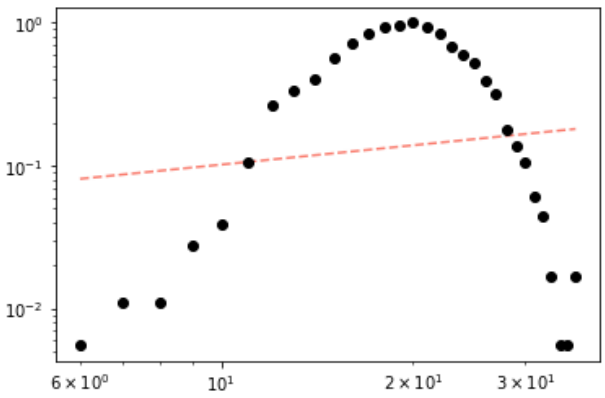




GNP2

This graph is not scale free as it does not follow power law.

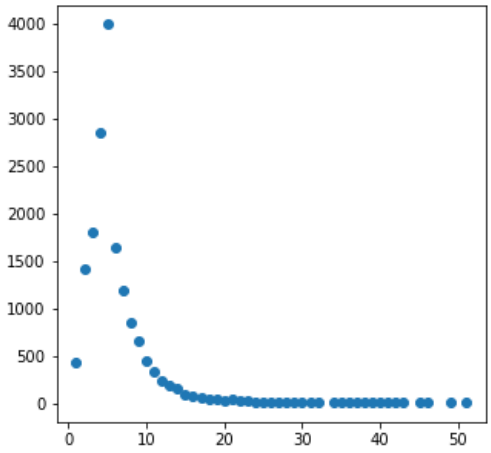


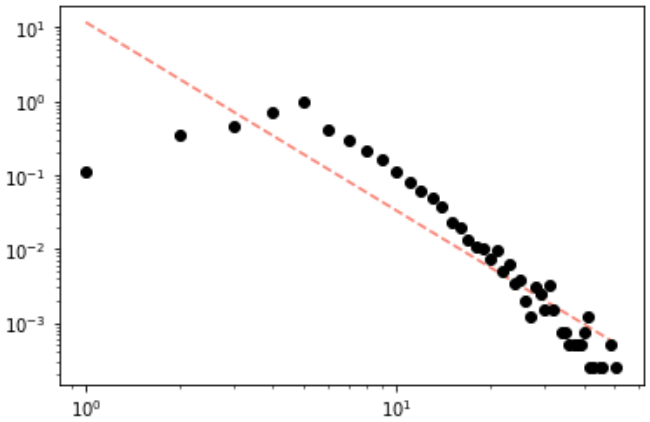


2. Do the Stanford graphs provided to you appear to be scale free?

Amazon Graph Large

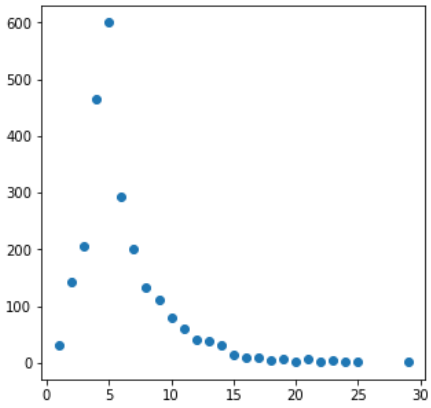
This graph is scale free as it does follow power law.

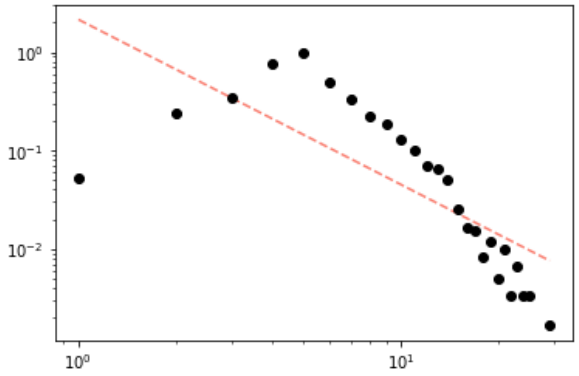




Amazon Graph Small

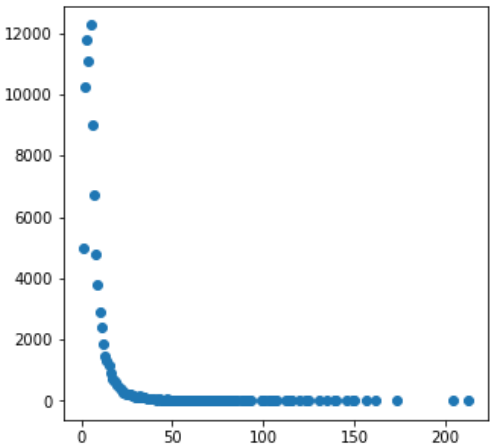
This graph is scale free as it does follow power law.

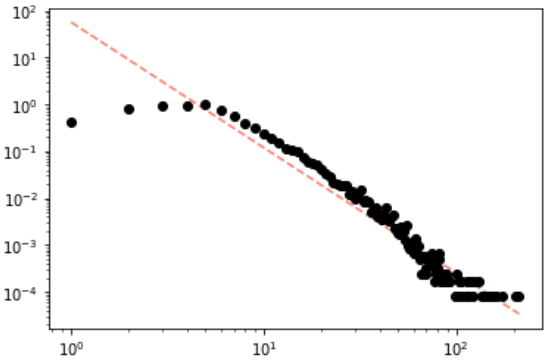




DBLP Graph Large

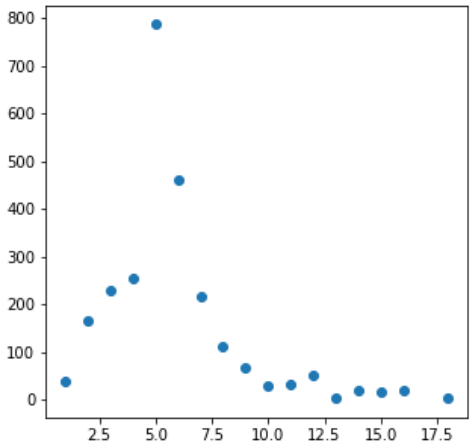
This graph is scale free as it does follow power law.

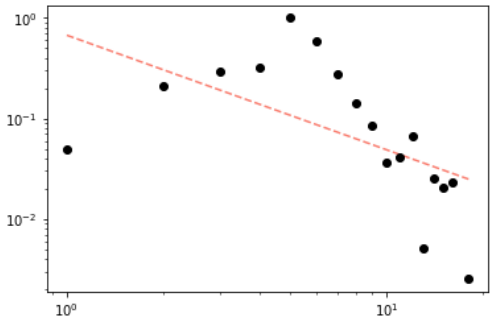




DBLP Graph Small

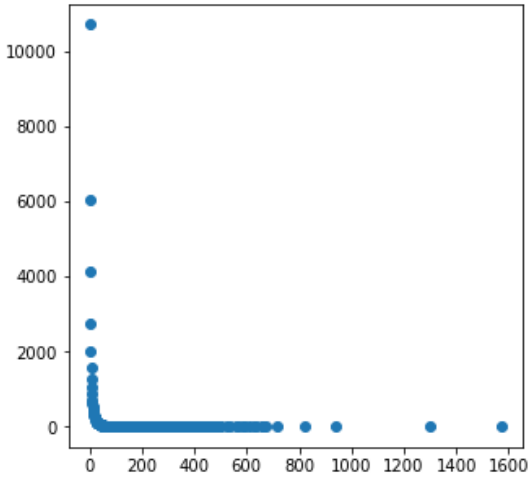
This graph is not scale free as it does not follow power law. But above 5, it does seem to follow power law.

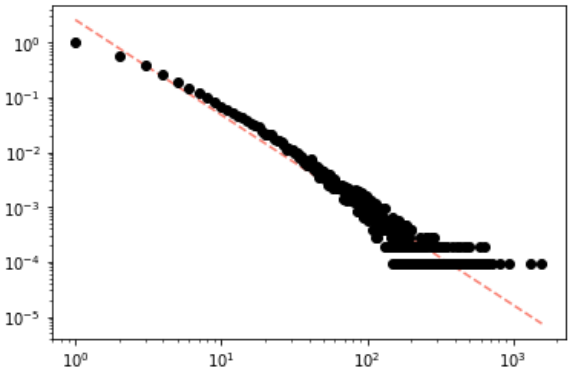




YouTube Graph Large

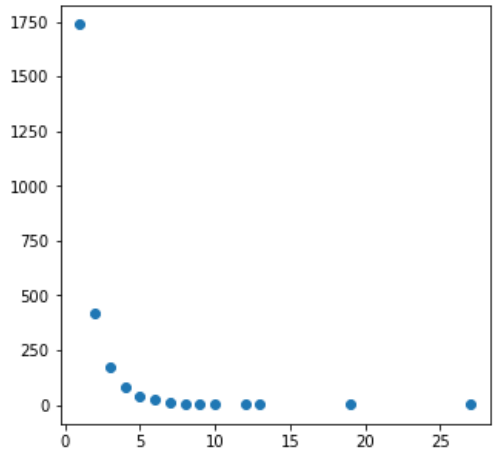
This graph is scale free as it does follow power law.

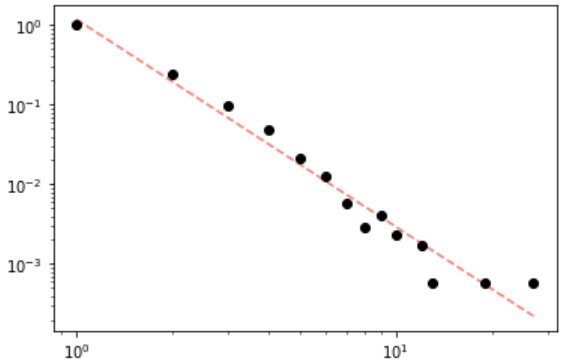




YouTube Graph Small

This graph is scale free as it does follow power law.





Centrality

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

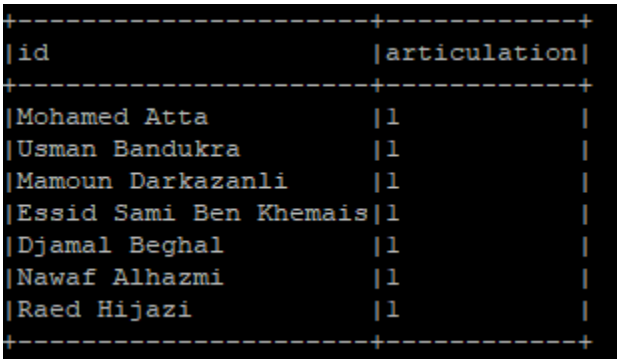
```
+---+--------------------+
| id|           closeness|
+---+--------------------+
|  F| 0.07142857142857142|
|  C| 0.07142857142857142|
|  H| 0.06666666666666667|
|  D| 0.06666666666666667|
|  B|0.058823529411764705|
|  E|0.058823529411764705|
|  G| 0.05555555555555555|
|  A| 0.05555555555555555|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

rspillai@vm16-239:/afs/unity.r
```

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: Machines with the highest closeness will be the best candidates. In our case it is machine F and C. Their sum of shortest paths is the least and thus, the number of hops required will also be the least.

Articulation Points

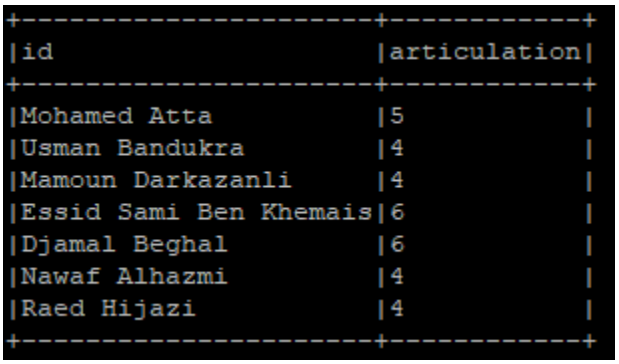
1. In this example, which members should have been targeted to best disrupt communication in the organization?
   Answer:

Processing graph using Spark iteration over nodes and serial (networkx) connectedness calculations.



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

Processing graph using serial iteration over nodes and GraphFrame connectedness calculations.



id	articulation
Mohamed Atta	5
Usman Bandukra	4
Mamoun Darkazanli	4
Essid Sami Ben Khemais	6
Djamal Beghal	6
Nawaf Alhazmi	4
Raed Hijazi	4

Essid Sami Ben Khemais, Djamal Beghal and Mohamed Atta have the highest articulations. Thus, if they are targeted, this will disrupt the communication in the most effective way.