CSC 591 Algorithms for Data Guided Business Intelligence

Project - Network Properties in Spark GraphFrames

Krishna Murali
kmurali2

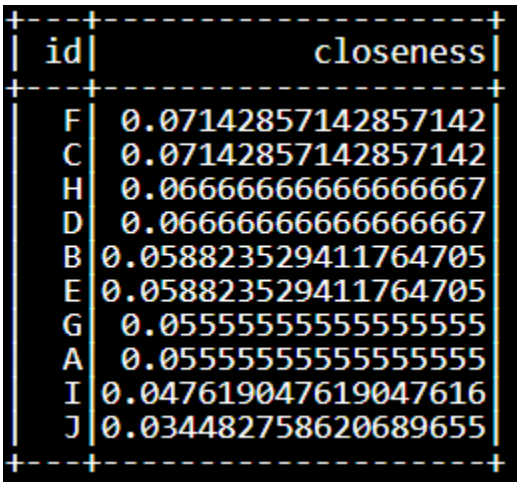
Degree Distribution:

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) (Powerlaw package is required)
   a. **gnm1**
      - **γ = 2.88754** As γ is between 2 and 3, gnm1 is **scale free**
   b. **gnm2**
      - **γ = 9.62066** As γ is greater than 3, gnm2 is **not scale free**
   c. **gnp2**
      - **γ = 54.58226** As γ is greater than 3, gnp2 is **not scale free**
   d. **gnp1**
      - **γ = 4.93908** As γ is greater than 3, gnp1 is **not scale free**

2. Do the Stanford graphs provided to you appear to be scale free?
   a. **amazon.graph.large**
      - **γ = 1.325577** As γ is lesser than 3, amazon.graph.large is **not scale free**
   b. **amazon.graph.small**
      - **γ = 2.39486** As γ is between 2 and 3, amazon.graph.small is **scale free**
   c. **youtube.graph.small**
      - **γ = 1.36744** As γ is lesser than 3, youtube.graph.small is **not scale free**
   d. **youtube.graph.large**
      - **γ = 1.56051** As γ is lesser than 3, youtube.graph.large is **not scale free**
   e. **dblp.graph.large**
      - **γ = 1.31439** As γ is lesser than 3, dblp.graph.large is **not scale free**
   f. **dblp.graph.small**
      - **γ = 1.60778** As γ is lesser than 3, dblp.graph.large is **not scale free**

Centrality:

1. Rank the nodes from highest to lowest closeness centrality



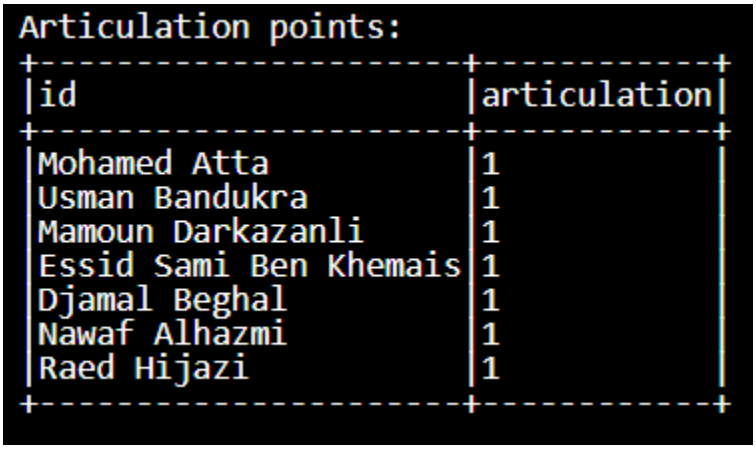
id	closeness
F	0.07142857142857142
C	0.07142857142857142
H	0.06666666666666667
D	0.06666666666666667
B	0.058823529411764705
E	0.058823529411764705
G	0.055555555555555555
A	0.055555555555555555
I	0.047619047619047616
J	0.034482758620689655

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?

   Depending on the closeness values, Nodes **F and C** will be the best candidates to hold the centralized data.

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