Web and Social Computing (IT 752)

Submitted by: Sampat Kr Ghosh
Roll Number: 192IT020
Submitted to: Dr. Sowmya Kamath S.

Lab Assignment 3

Dataset

1. Facebook
2. Gnutella P2P Network, August 4, 2002
3. Twitter

Packages Used

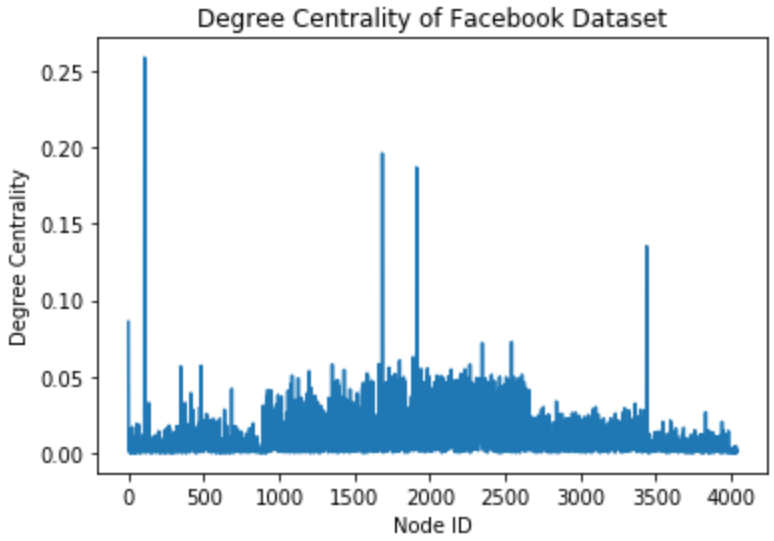
1. NetworkX
2. Matplotlib
3. Numpy
4. Pandas

Part 1

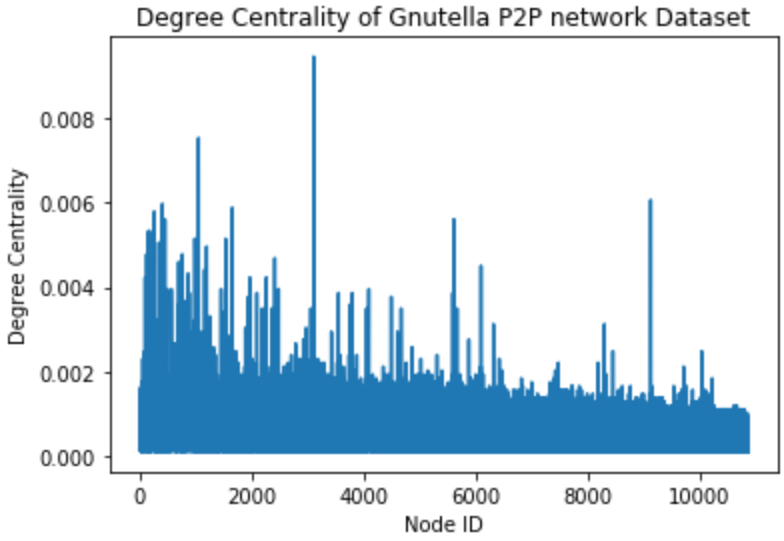
The datasets mentioned above were used to measure the centrality of nodes - Degree centrality, Closeness centrality, Betweenness centrality, and Eigenvector centrality. For plotting the distributions, normal and logarithmic scale both are used as for some distributions that are in normal scale, I was not getting a good plot.

Degree Centrality

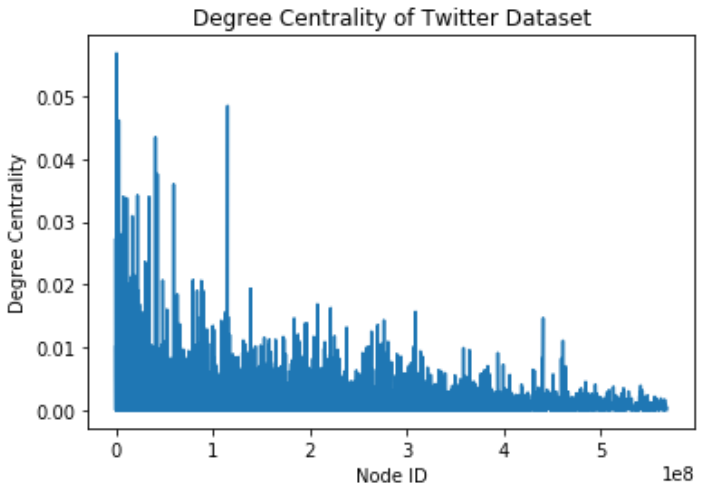
Degree centrality is the simplest centrality measure to compute. Recall that a node's degree is simply a count of how many social connections (i.e., edges) it has. The degree centrality for a node is simply its degree. For degree centrality, higher values mean that the node is more central. Degree centrality shows how many connections a person has. They may be connected to lots of people at the heart of the network, but they might also be far off on the edge of the network.
Score (C_d) = number of edges attached to the node.
Standardized Score = C_d / (n-1), where n is the total number of nodes.
Degree Centrality V/S Node ID plots are given below:



Facebook

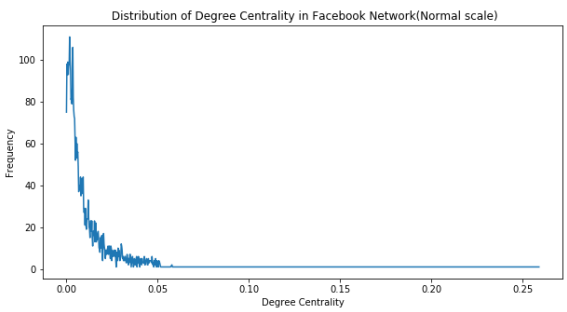


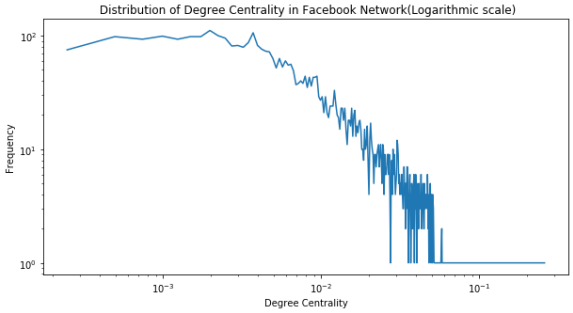
Gnutella P2P Network



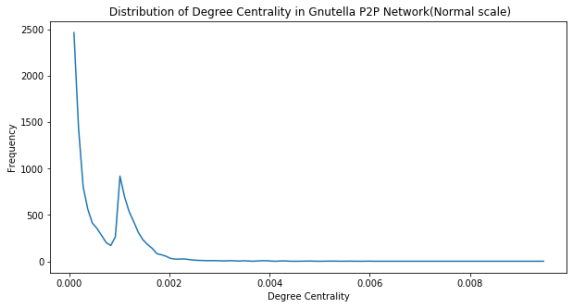
Twitter

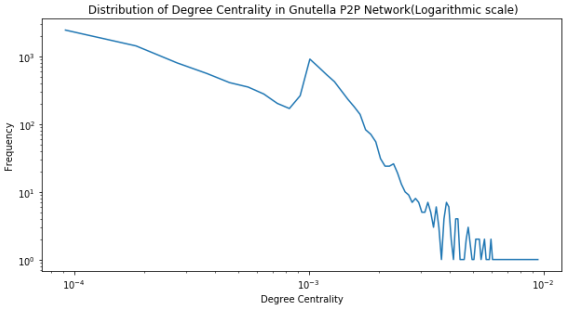
Degree centrality distribution for all the datasets are plotted below:



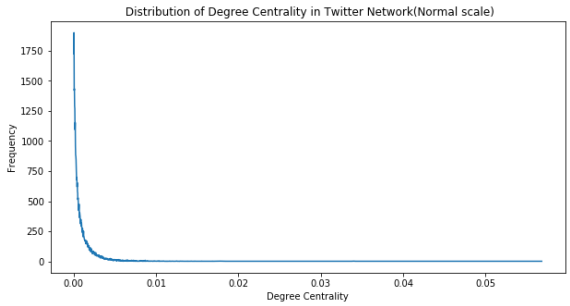


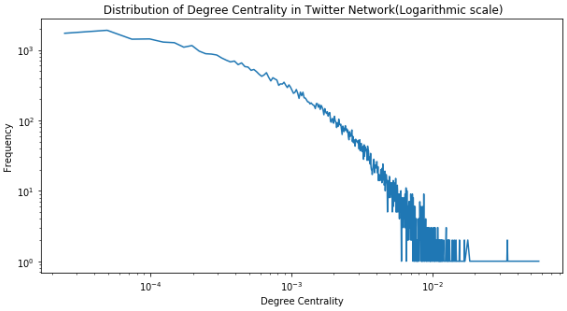
Facebook





Gnutella P2P Network





Twitter

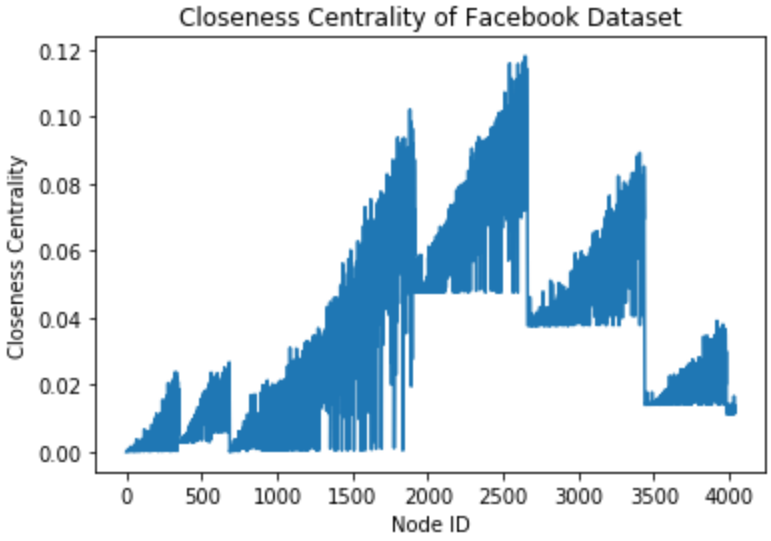
Observation

Similar to Degree distribution the degree centrality distribution also shows power law of distribution ie. very few nodes have higher degree centrality and the majority of nodes have very small degree centrality.

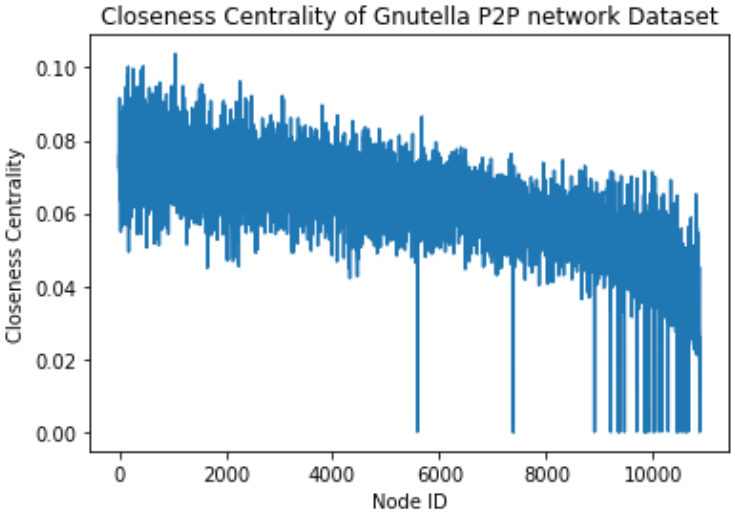
Closeness Centrality

Closeness centrality indicates how close a node is to all other nodes in the network. It is calculated as the average of the shortest path length from the node to every other node in the network. It is defined as the reciprocal of the average shortest path length.
Score = 1 / avg (L(n,m)), where L(n,m) is the length of the shortest path between two nodes n and m.

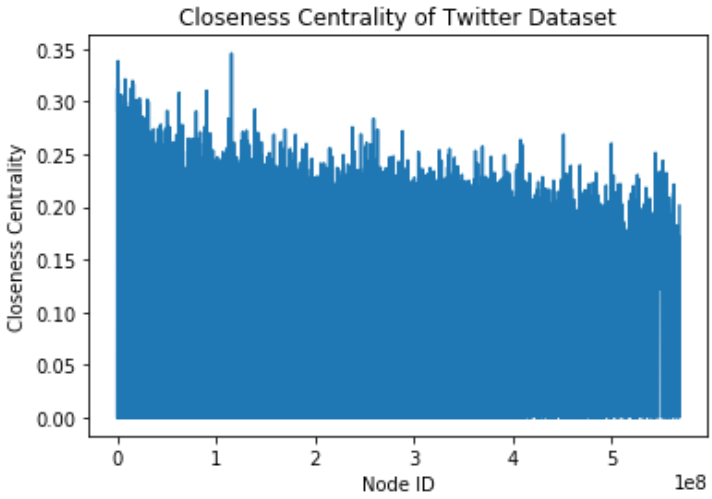
Closeness centrality V/S Node ID plots are given below:



Facebook

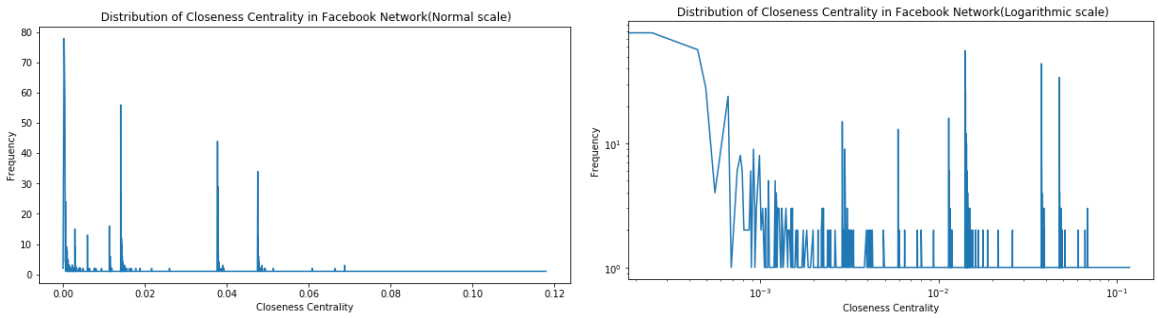


Gnutella P2P Network

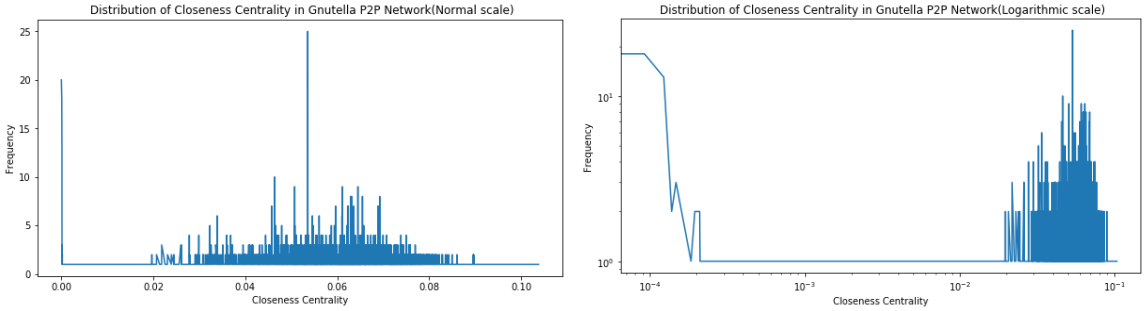


Twitter

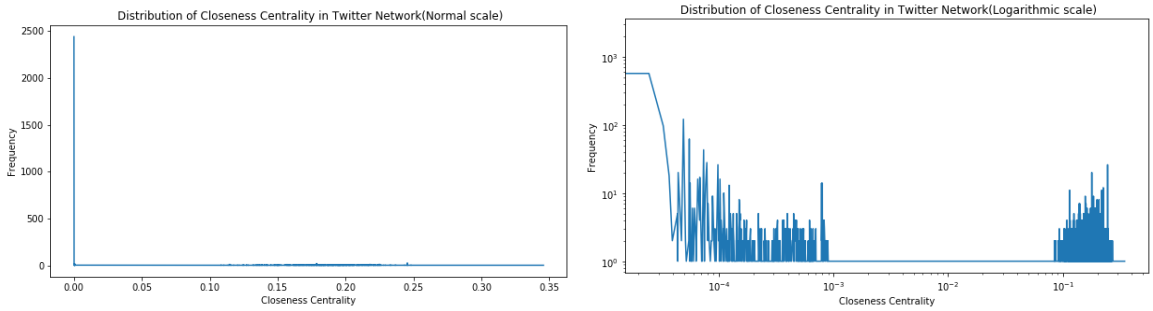
Closeness centrality distribution for all the datasets are plotted below:



Facebook



Gnutella P2P Network



Twitter

Observation

Closeness centrality can help find good 'broadcasters', but in a highly-connected network, you will often find all nodes have a similar score. What may be more useful is using Closeness to find influencers in a single cluster. As the above datasets are not highly connected, we have dissimilar closeness centrality score.

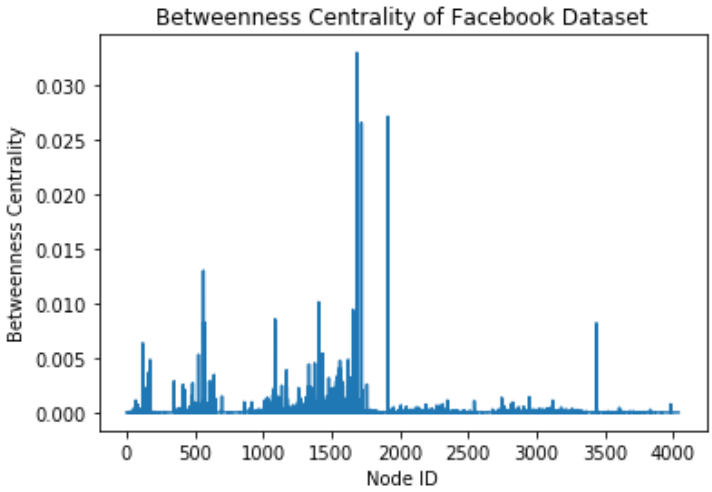
Betweenness Centrality

Betweenness centrality measures how important a node is to the shortest paths through the network. To compute betweenness for a node N, we select a pair of nodes and find all the shortest paths between those nodes. Then we compute the fraction of those shortest paths that include node N.

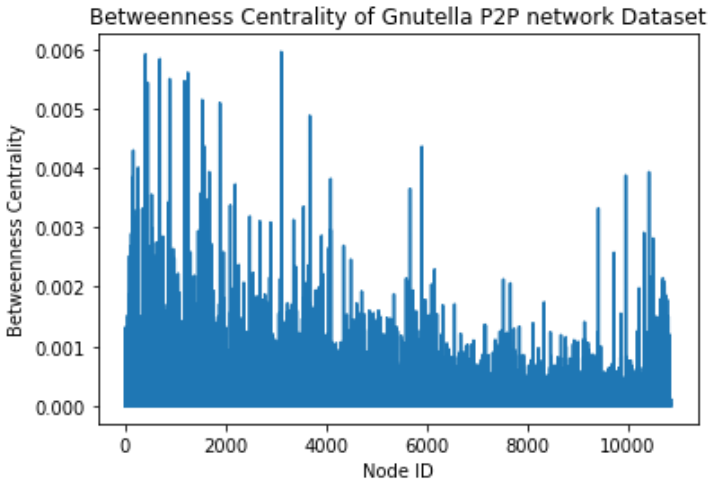
Score $(C_b) = \Sigma\sigma_{st}(v)/\sigma_{st}$, where $\sigma_{st}(v)$ is the number of shortest paths from s to t that v lies on and σ_{st} is the number of shortest paths from s to t.

Standardized Score = $C_b/[(n-1)(n-2)/2]$

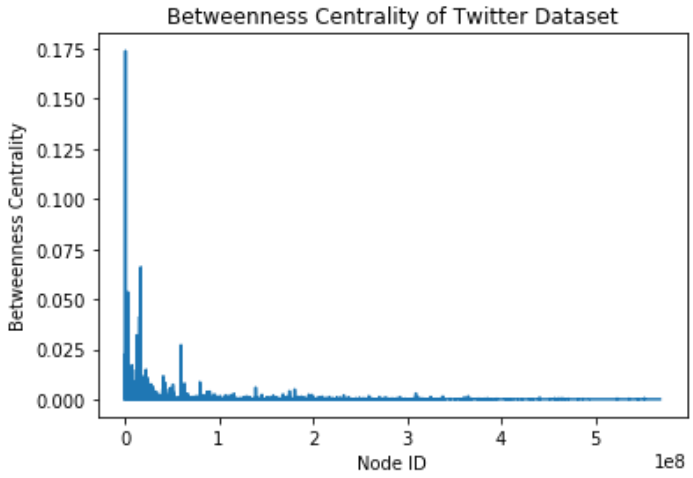
Betweenness centrality V/S Node ID plots are given below:



Facebook

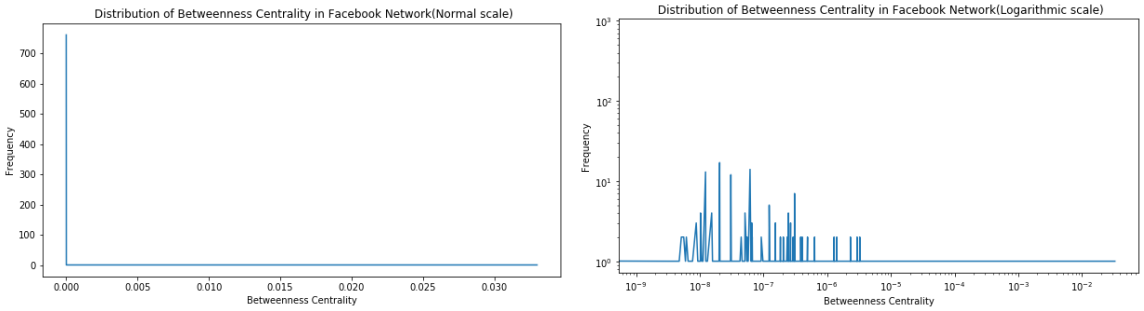


Gnutella P2P Network

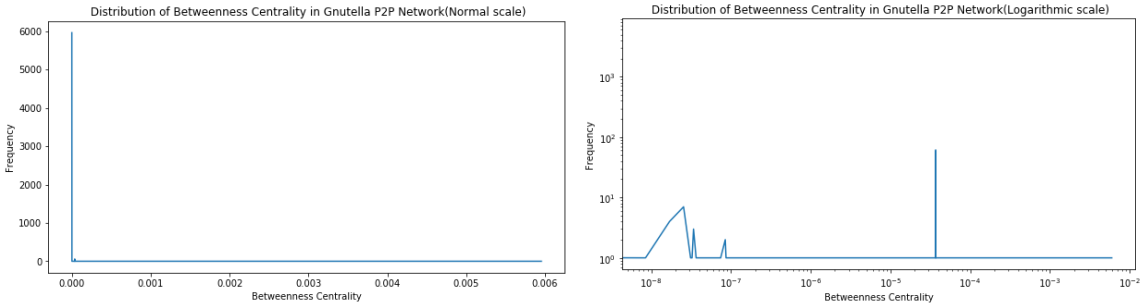


Twitter

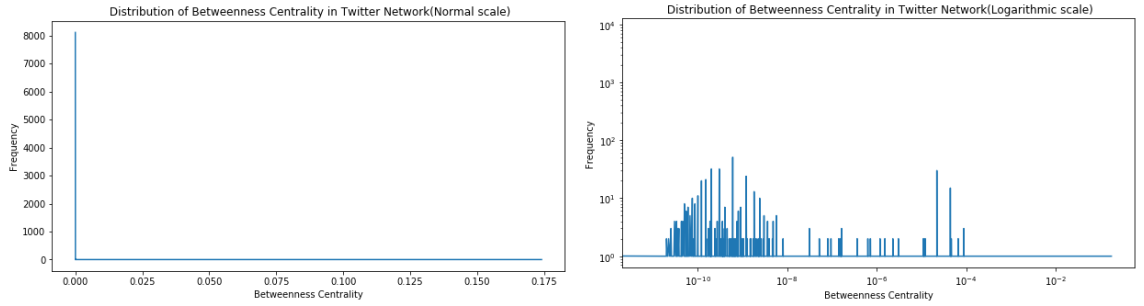
Betweenness centrality distribution for all the datasets are plotted below:



Facebook



Gnutella P2P Network



Twitter

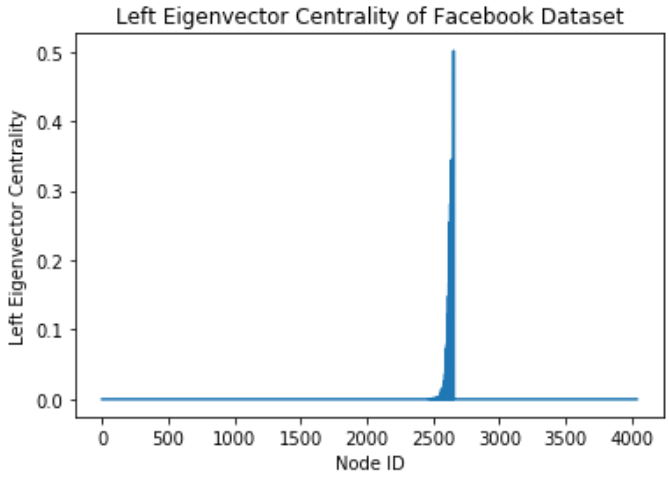
Observation

A high betweenness centrality could indicate someone holds authority over disparate clusters in a network, or just that they are on the periphery of both clusters. As we can see from the above plots, there are few nodes that have high betweenness centrality. From this, we can conclude that these individuals influence the flow around a system.

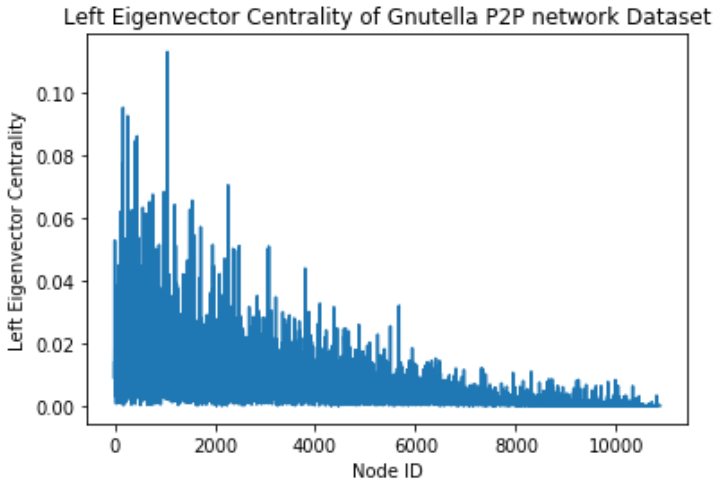
Eigenvector Centrality

Eigenvector centrality is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Google's PageRank and the Katz centrality are variants of the eigenvector centrality. Since the datasets used here are directed graphs,

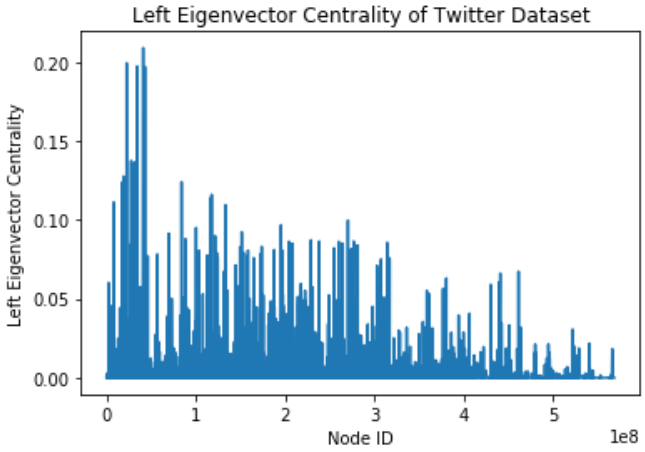
we need to compute Left Eigenvector (Incoming edge) Centrality and Right Eigenvector Centrality (outgoing edges).
Left eigenvector centrality V/S Node ID plots are given below:



Facebook

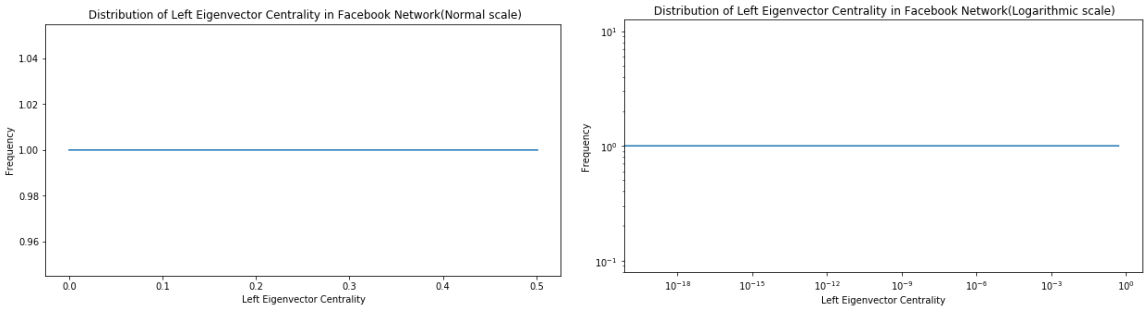


Gnutella P2P Network

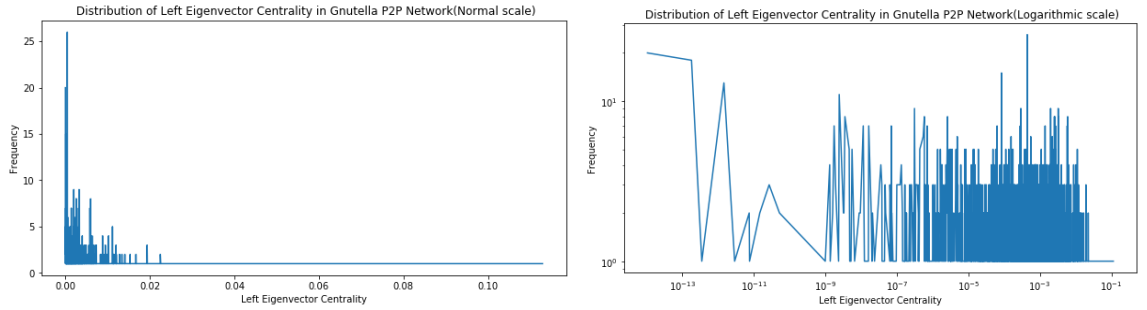


Twitter

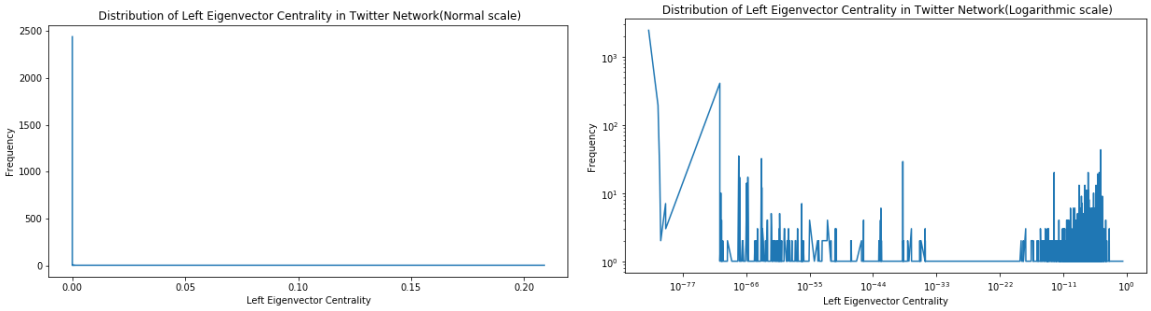
Left eigenvector centrality distribution for all the datasets are plotted below:



Facebook

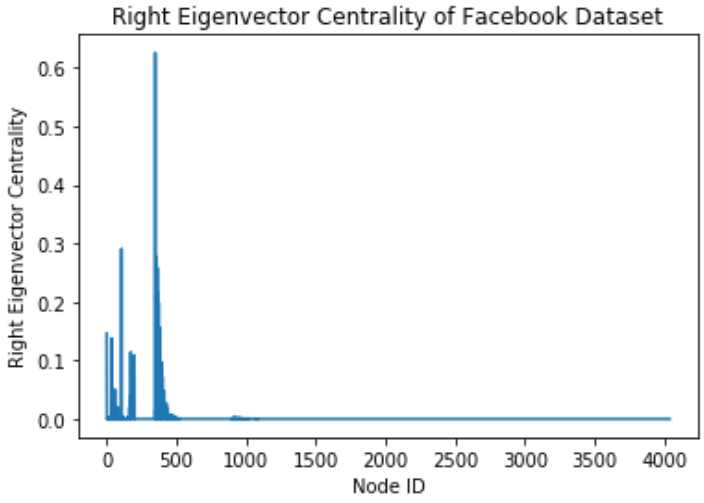


Gnutella P2P Network

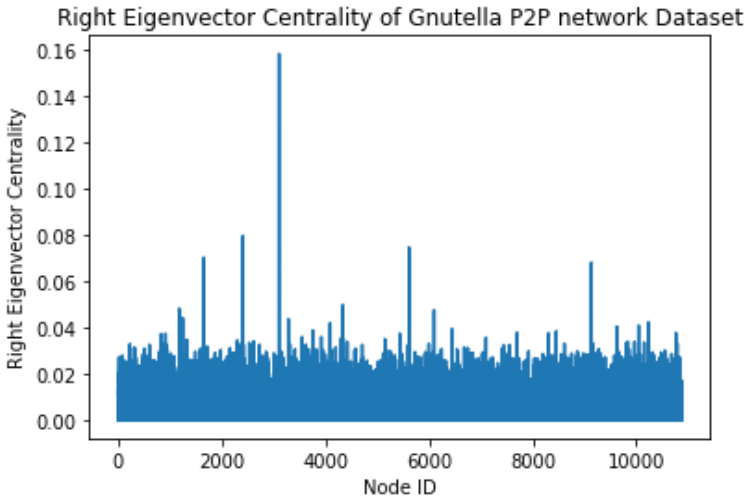


Twitter

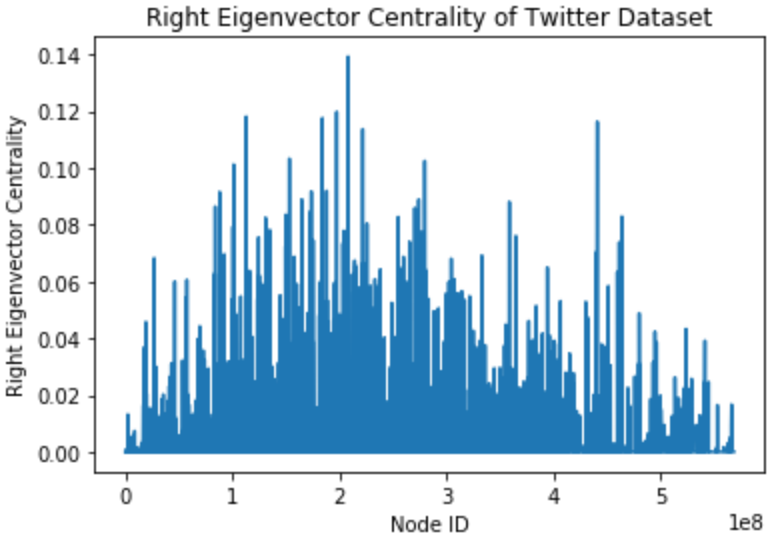
Right eigenvector centrality V/S Node ID plots are given below:



Facebook

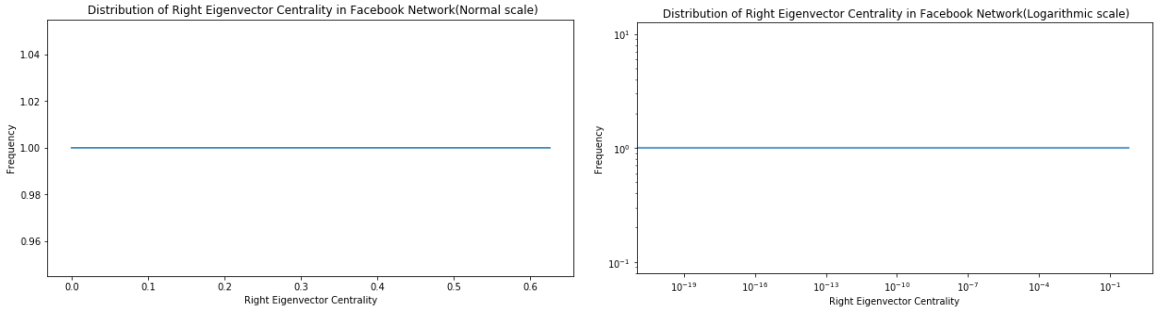


Gnutella P2P Network

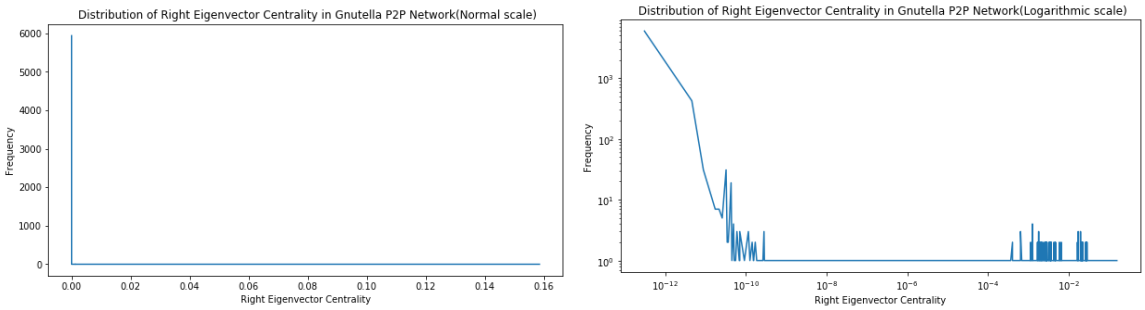


Twitter

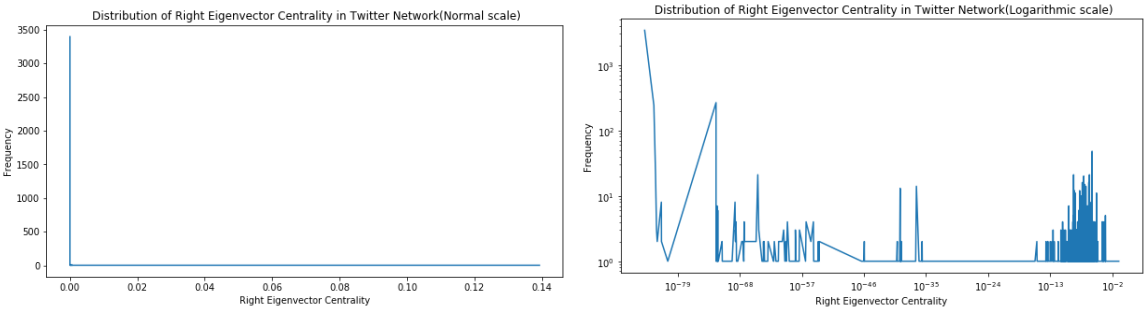
Right eigenvector centrality distribution for all the datasets are plotted below:



Facebook



Gnutella P2P Network



Twitter

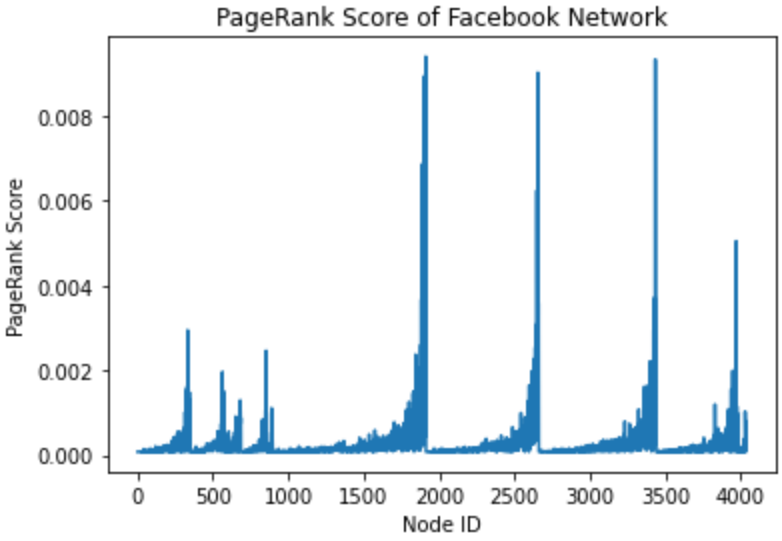
Observation

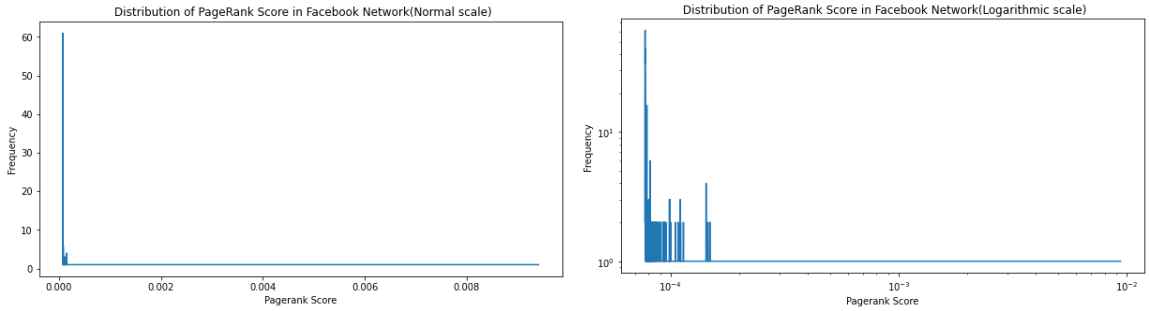
By calculating the extended connections of a node, Eigenvector Centrality can identify nodes with influence over the whole network, not just those directly connected to it. As we know that eigenvector centrality is a measure of the influence of a node in a network, from the above plots we can observe that very few nodes have high centrality and large numbers of nodes have low centrality. To interpret social networking websites, the nodes with high centrality are potential influencers.

Part 2

The datasets mentioned above were used to implement and compare PageRank and HITS Score algorithms. For plotting the distributions, normal and logarithmic scale both are used as for some distributions that are in normal scale, I was not getting a good plot.

Facebook Dataset





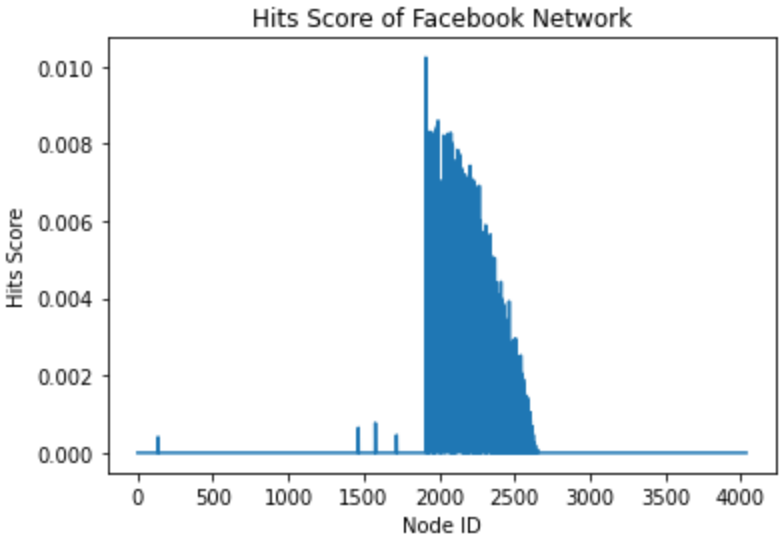
PageRank

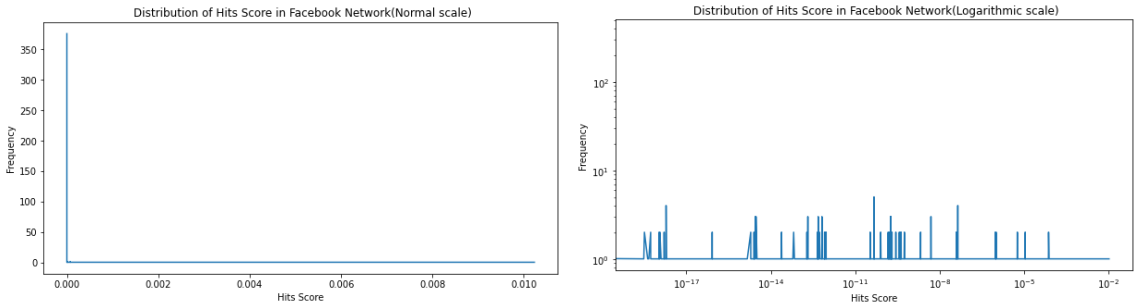
Max PageRank Score = 0.00940916451585816

Min PageRank Score = 7.724494426592372e-05

Observation

From the plot it is clear that only a few nodes have a very high PageRank score compared to other nodes that form the majority of the network. This ensures the trend of social networks where there are much fewer highly influential users like celebrities and a high number of normal users with low influence in the network. From the distribution plot we can see that it is a highly skewed Power-Law distribution. I have plotted both in logarithmic and normal scale since the values are very small and differ by a very small margin.

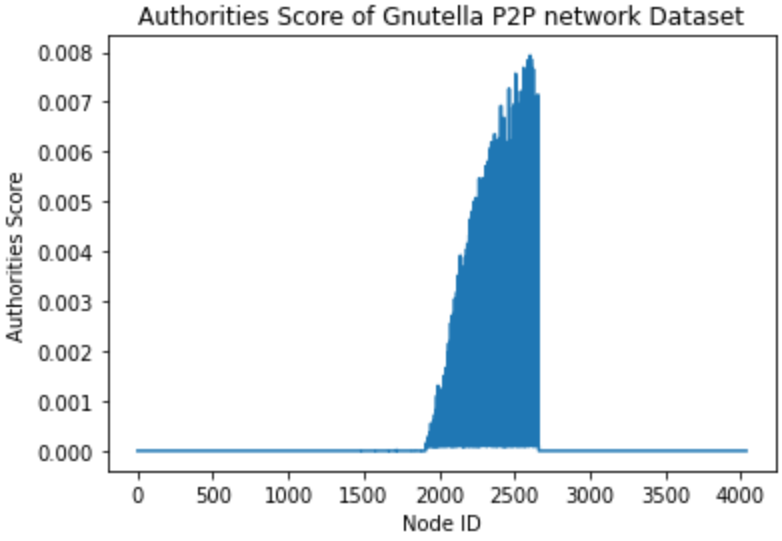


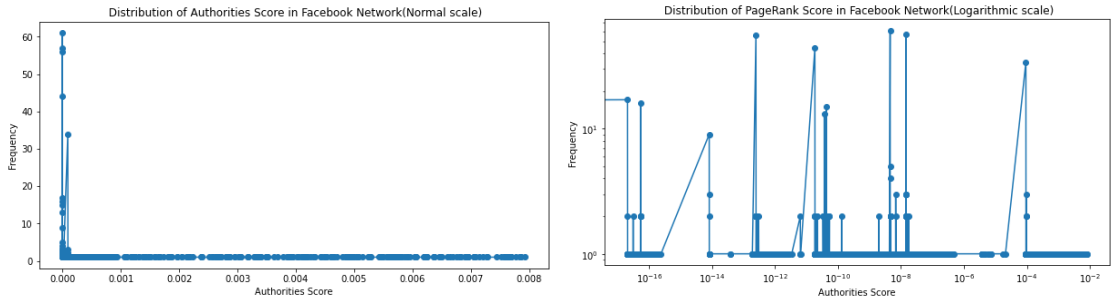


Hits Score

Max Hits Score = 0.010229403947448484

Min Hits Score = 0.0





Authorities Score

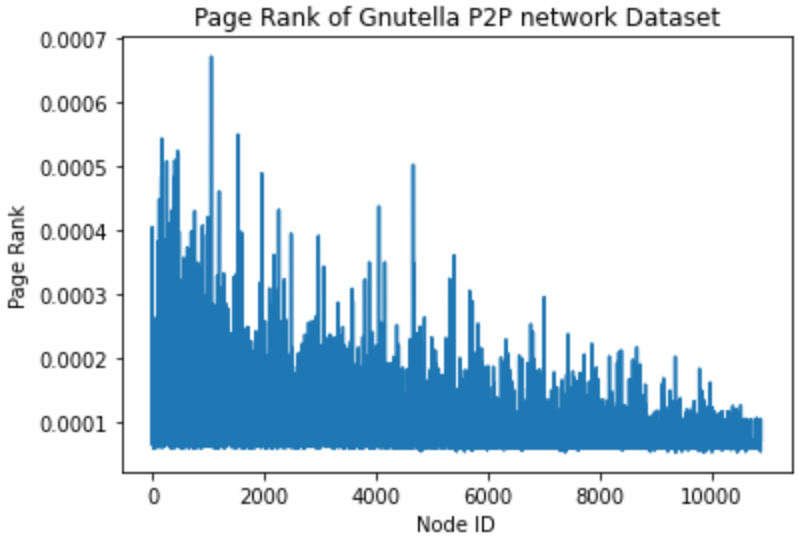
Max Authorities Score = 0.007932133892198798
Min Authorities Score = 0.0

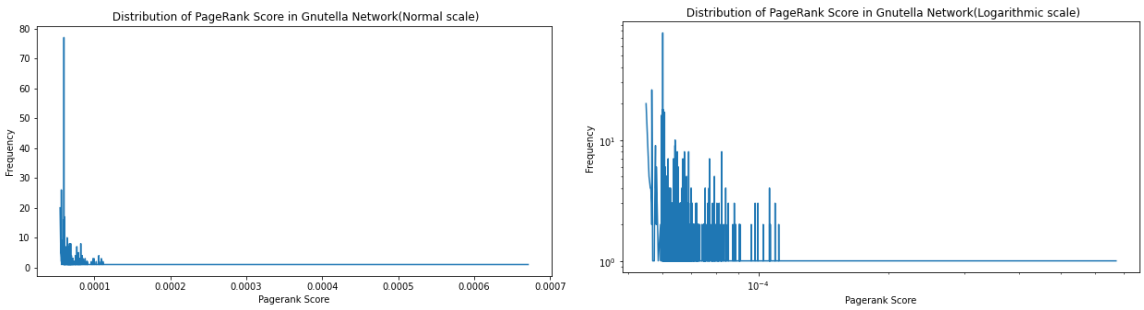
Observation

HITS and Authorities score shows a similar trend as it was in Pagerank. Only a few nodes have high HITs and Authorities score while most of the nodes are very low. The difference is the high PageRank score is more distributed among the node identifiers while nodes with high HITs and Authorities score are clustered together. From the normal scale plot it is clear that HITS algorithm gives Power-Law distribution. The logarithm scale plot is only made because the values are too small to have proper visuals in the normal scale plot.

We have also observed that Maximum and minimum Pagerank, Hits and Authorities score similar and close to each other. This shows these two algorithms measure the similar properties of nodes in a network and in a similar scale.

Gnutella P2P Network



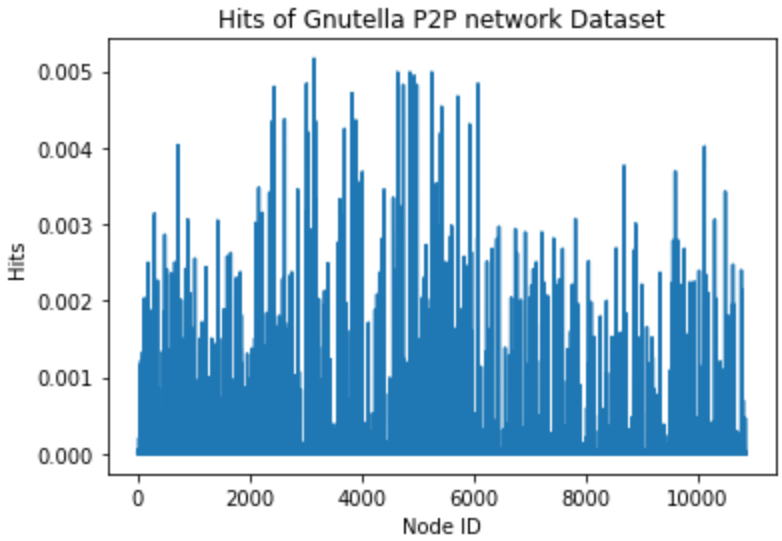


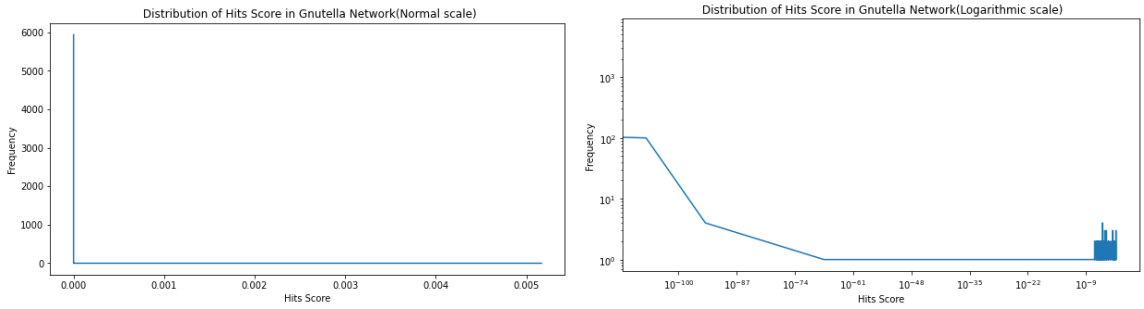
PageRank

Max PageRank Score = 0.0006711727183638689
Min PageRank Score = 5.499573860784478e-05

Observation

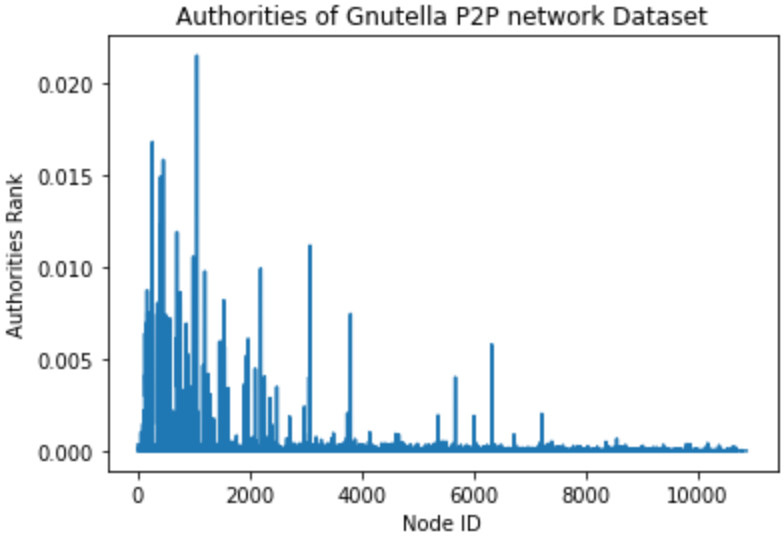
Observations are the same as that of the Facebook dataset.

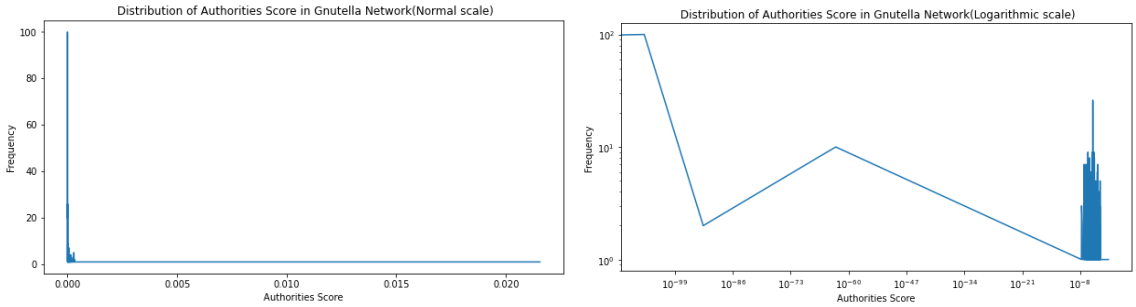




Hits Score

Max Hits Score = 0.005167046979475104
Min Hits Score = 0.0





Authorities Score

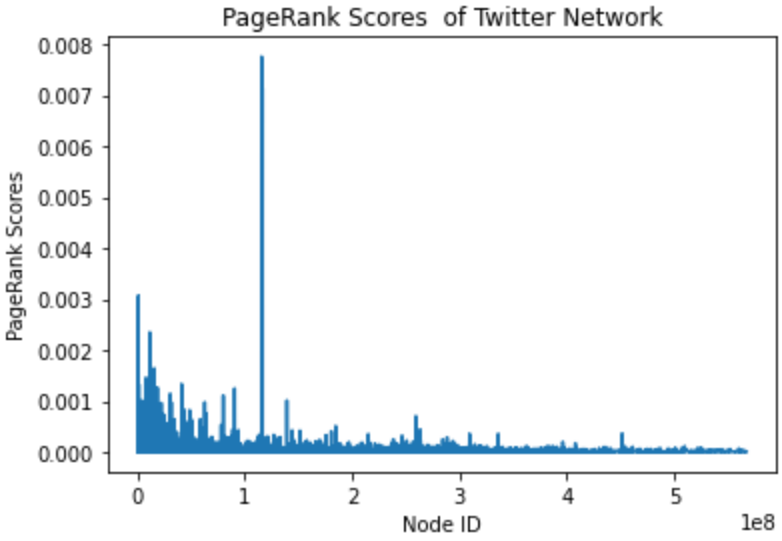
Max Authorities Score = 0.021553778629464077

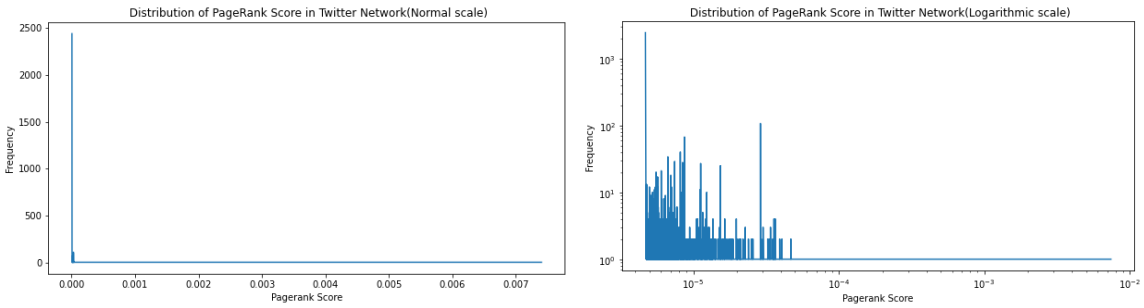
Min Authorities Score = 0.0

Observation

Observations are the same as that of the Facebook dataset.

Twitter Dataset



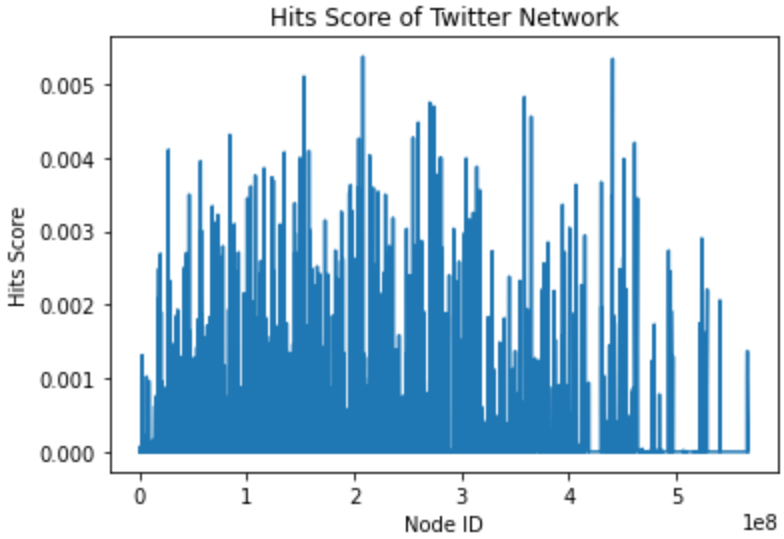


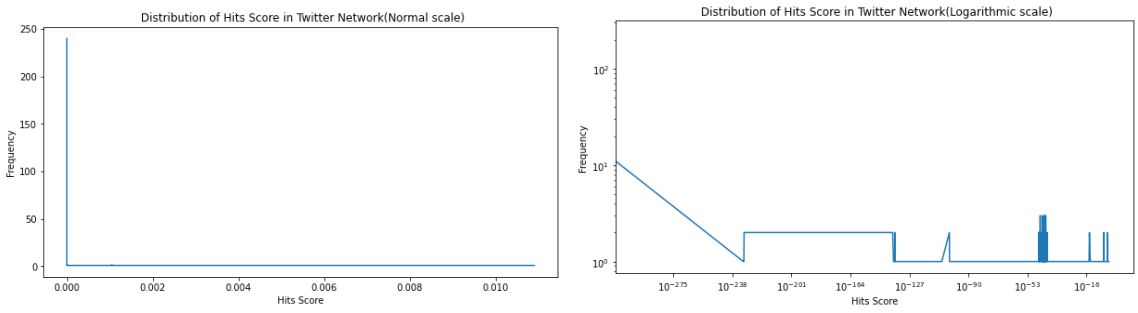
PageRank

Max PageRank Score = 0.007408496105825751
Min PageRank Score = 4.689332128677982e-06

Observation

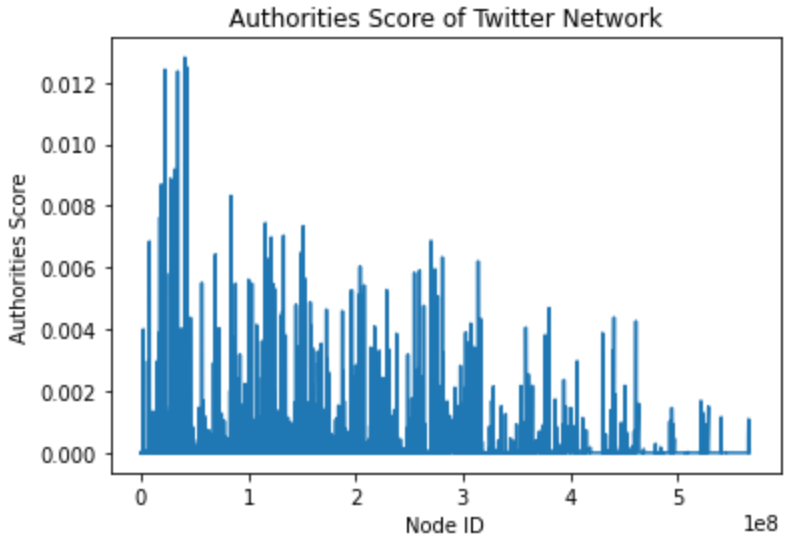
Observations are the same as that of the Facebook dataset.

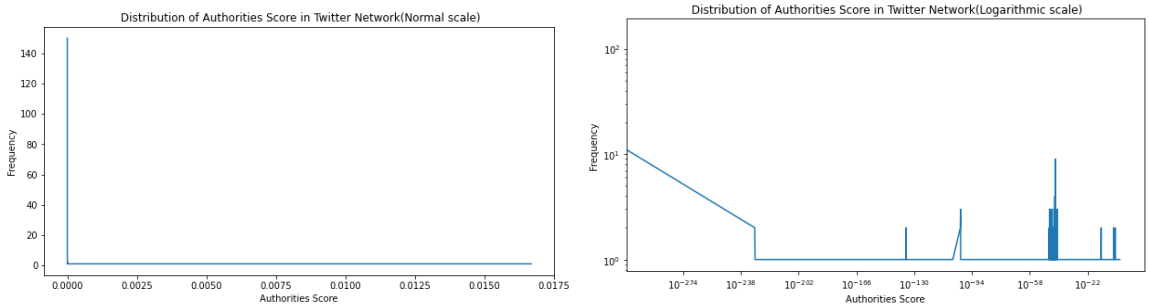




Hits Score

Max Hits Score = 0.010907823237657264
Min Hits Score = 0.0





Authorities Score

Max Authorities Score = 0.01668608174631693
Min Authorities Score = 0.0

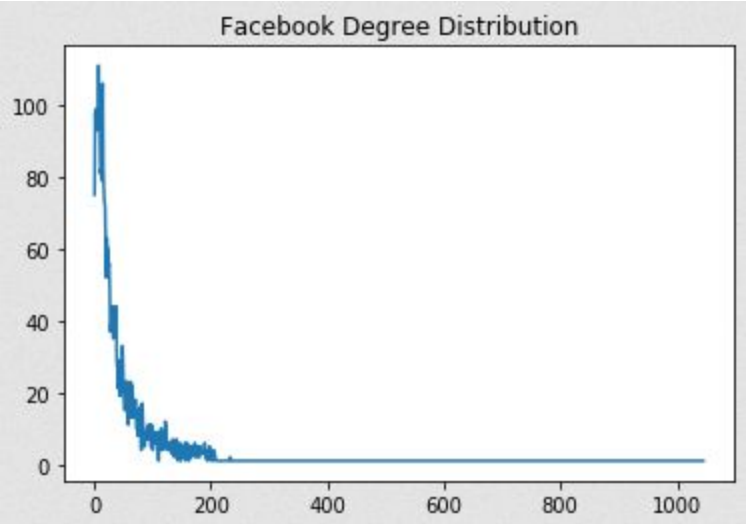
Observation

Observations are the same as that of the Facebook dataset.

Correlation of PageRank/HITS and fundamental network properties:

One of the fundamental network properties is the degree of a node. Degree of a node is a parameter of how many nodes are connected to that node. So if we can plot the distribution of degrees of the nodes in a network it can capture how many nodes have higher degrees hence more connected and how many have low degrees hence less connected. PageRank and HITS similar properties are calculated for a node in network to assess the relevancy of that node. The higher connected the node is, the better PageRank score or HITS score it has.

Previously I calculated and plotted the degree distribution of the Facebook network dataset and now with Pagerank and HITS calculated for the same network I found similarities between the distribution of both the assessments. The following plot shows the degree distribution of the data set. It is clear that a high number of nodes have a very small degree and a very low number of nodes shows a high degree.



From the PageRank and HITS score distribution from above have confirmed the same property where a very low number of nodes had high scores. Also, such a network has a very low density. I observed in assignment 1 that the Facebook dataset had a density = 0.0108199635. This is very low compared to the highest value density can have, which is 1. This is because nodes in social networks are not very well connected and many nodes have much fewer edges. Both score distribution and degree distribution supports this property.

Conclusion:

From the observations and analysis from above it can be concluded that from the score of nodes with algorithms like HITs and PageRank we can draw many properties of the network under study. These properties are also supported by the observations of fundamental network analysis like degree distribution, density, etc. So, these are correlated. This analysis on the Facebook dataset confirms a trend in social media networks as well as many such networks and using any of the techniques mentioned above individually or collectively we can thoroughly analyze a network.