ASSIGNMENT REPORT

On

CENTRALITY, HITS AND PAGERANK

Under the guidance of
Dr. Sowmya Kamath

Submitted By :

PRAJWAL M P(192IT015)

PURPOSE

[1] To implement and measure various normalized metrics used to measure centrality of nodes - Degree centrality, Closeness centrality, Betweenness centrality and Eigenvector centrality.

[2] To implement (i) HITS algorithm (ii) PageRank algorithm and compare them.

INTRODUCTION

Centrality tries to answer the question as to what makes a node important. The centrality is given in terms of a real valued function on the nodes of a graph G, where the values produced provide a ranking which identifies the most important nodes.

There are various methods of ranking a vertex or a node in a graph to identify how important they are in a network or graph. Some of them are

- **Degree Centrality**: Degree centrality is the count of edges onto a node or simply it is the degree of that vertex. Since more connected nodes tend to have a greater number of connections to them degree of a node gives us a crude way of identifying how important a node is. The degree centrality of a vertex v , for a given graph $G := (V, E)$ with $|V|$ vertices and $|E|$ edges, is defined as $C_D(v) = \deg(v)$.

- **Betweenness centrality**: Betweeness centrality is a centrality measure based on shortest paths. It is a count of number of shortest paths a given node is a part of. In other words, it is the number of shortest paths that pass through a node in a graph. The betweenness centrality of a node {\displaystyle v} v is given by the expression: $g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$

- **Closeness centrality**: Closeness centrality is a centrality measure which gives a higher ranking to node which closer to most of nodes or in other words is central to the network. It is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Closeness centrality of a node is given by $C(x) = \frac{1}{\sum_y d(y,x)}$.

- **Eigenvector Centrality**: Eigenvector Centrality is a centrality measure which assigns ranking based on influence of a node in a network. A node which is connected to other high scoring nodes gets a higher score than if it had been connected to low influence nodes. Some of the more famous web page ranking algorithms like pagerank and hits make use of this concept. For a given graph $G := (V, E)$ with $|V|$ vertices let $A = (a_{v,t})$ be the adjacency matrix, i.e. $a_{v,t} = 1$ if vertex v is linked to vertex t, and $a_{v,t} = 0$ otherwise. The relative centrality score of vertex v can be defined as: $x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$

WEB PAGE RANKING

Web page ranking algorithms are algorithms used mainly by search engines to rank web pages for a given query so as to produce satisfactory web links to the user. They try to identify relevant webpages based on score assigned to each site.

Some of the algorithms are [1] PageRank [2] HITS

PageRank : PageRank is a web page ranking algorithm used by google search engine to rank web pages in their search results. PageRank is a way of measuring the importance of website pages. According to Google:

"*PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.*"

The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called "iterations", through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

A probability is expressed as a numeric value between 0 and 1. A 0.5 probability is commonly expressed as a "50% chance" of something happening. Hence, a document with a PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to said document.

HITS: Hyperlink-Induced Topic Search (also known as hubs and authorities) is a link analysis algorithm that rates Web pages. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that they held, but were used as compilations of a broad catalog of information that led users direct to other authoritative pages. In other words, a good hub represents a page that pointed to many other pages, while a good authority represents a page that is linked by many different hubs

The Hub score and Authority score for a node is calculated with the following algorithm:

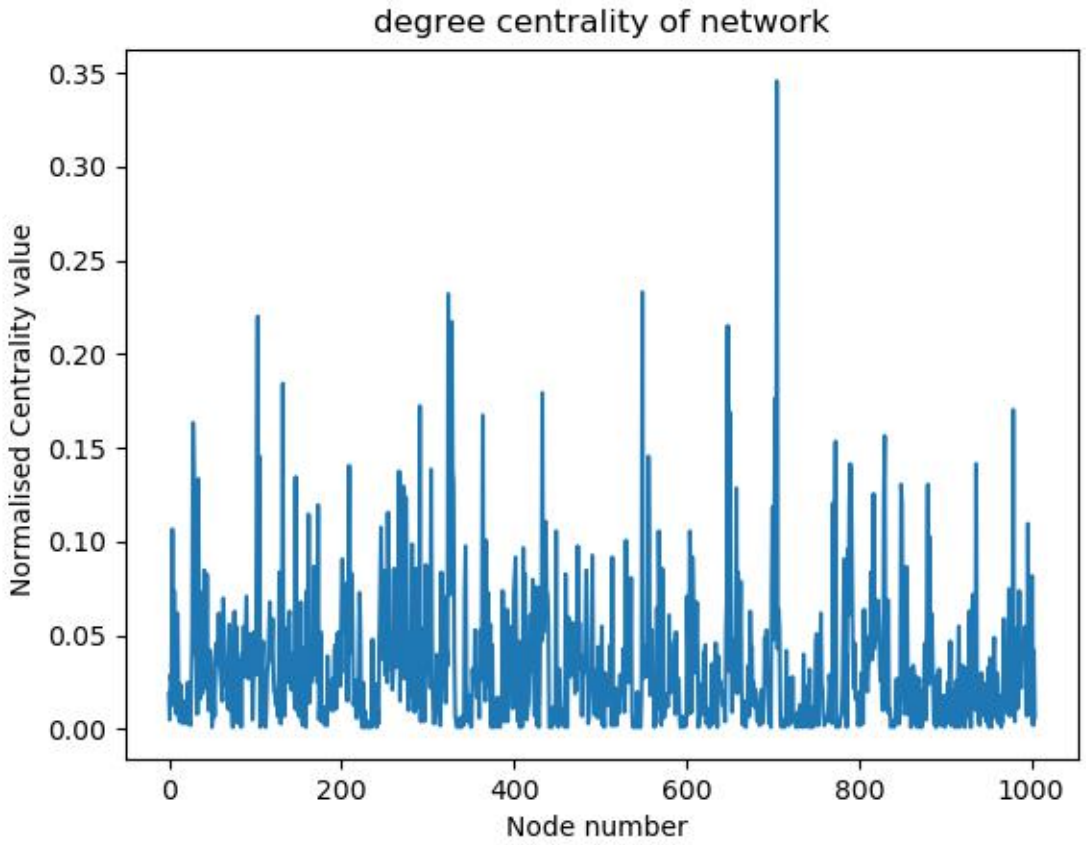
- Start with each node having a hub score and authority score of 1.
- Run the authority update rule
- Run the hub update rule
- Normalize the values by dividing each Hub score by square root of the sum of the squares of all Hub scores, and dividing each Authority score by square root of the sum of the squares of all Authority scores.
- Repeat from the second step as necessary

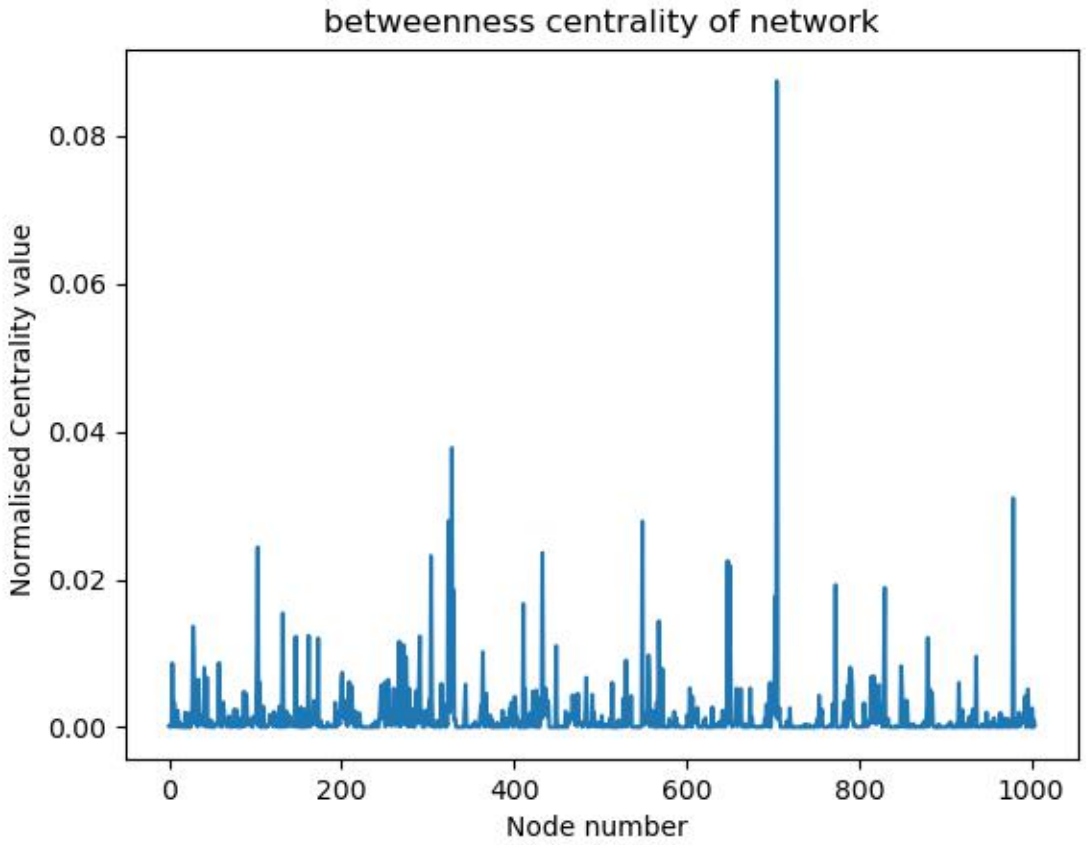
EXPERMENTATION AND RESULTS

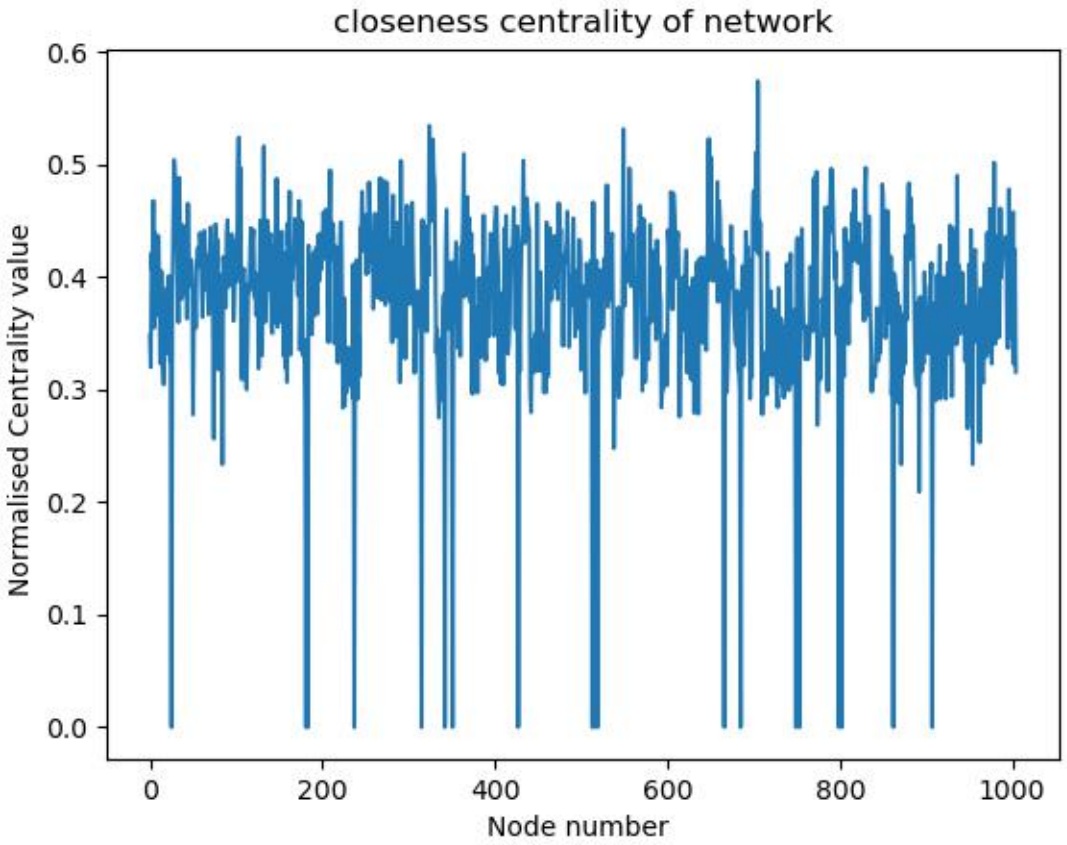
The centrality measures were calculated for four different datasets.

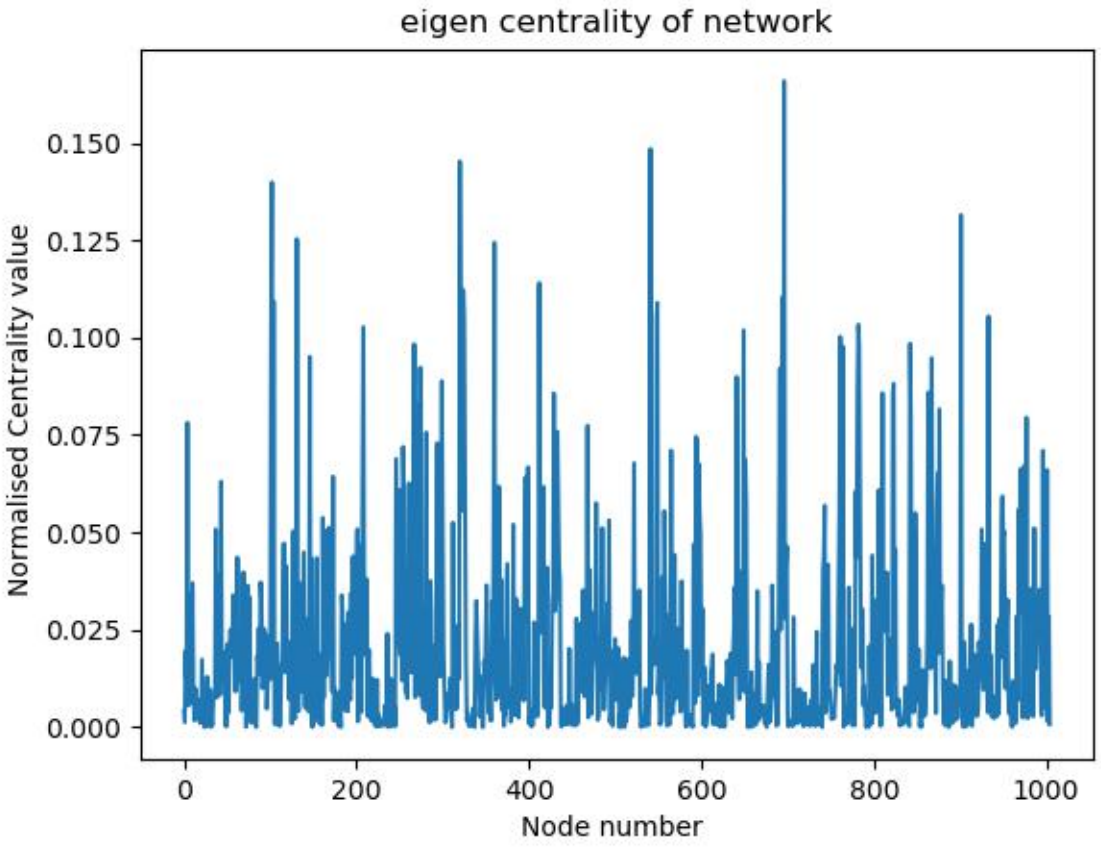
[1] email-Eu-core network

Dataset statistics	
Nodes	1005
Edges	25571
Nodes in largest WCC	986 (0.981)
Edges in largest WCC	25552 (0.999)
Nodes in largest SCC	803 (0.799)
Edges in largest SCC	24729 (0.967)
Average clustering coefficient	0.3994
Number of triangles	105461
Fraction of closed triangles	0.1085
Diameter (longest shortest path)	7
90-percentile effective diameter	2.9


degree centrality of network
Normalised Centrality value
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
200
400
600
800
1000
Node number


betweenness centrality of network
Normalised Centrality value
0.00
0.02
0.04
0.06
0.08
0
200
400
600
800
1000
Node number


closeness centrality of network
Normalised Centrality value
Node number
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
200
400
600
800
1000



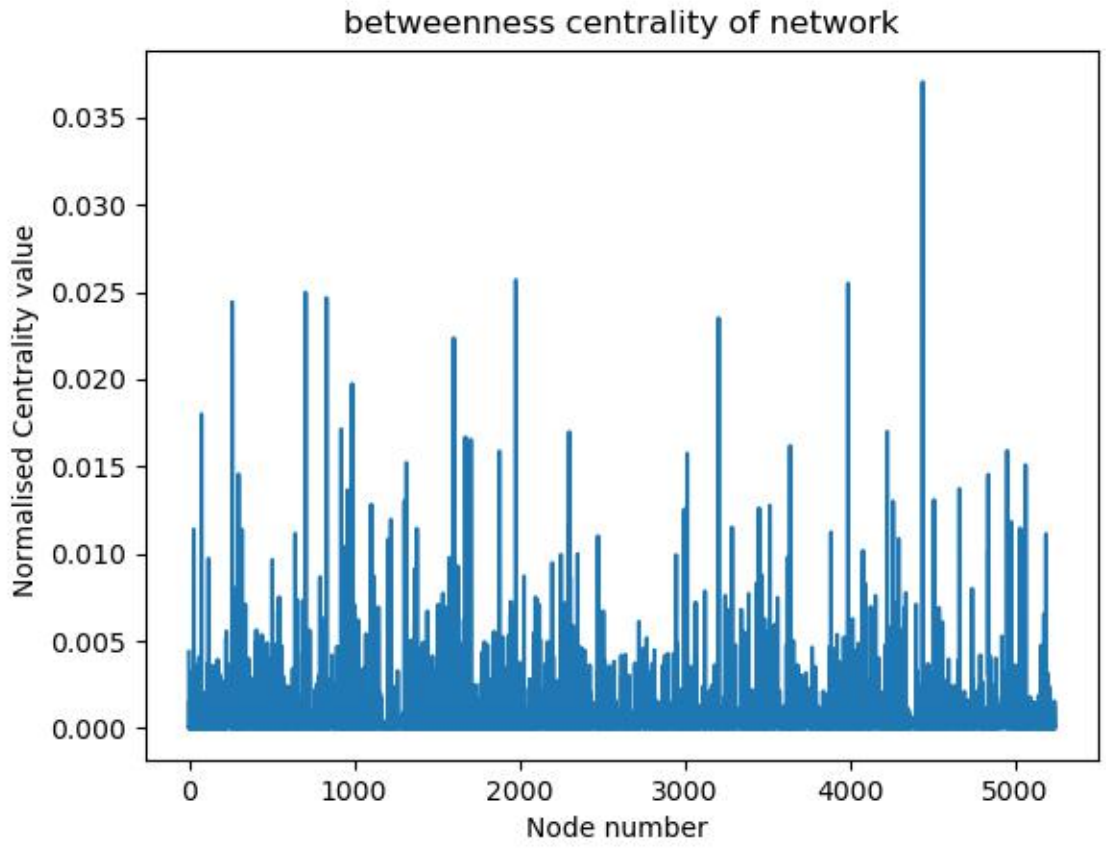
We can see that degree centrality, betweeness and eigen centrality are somewhat similar to these graphs. Since the diameter is very small we can see larger closeness centrality values also.

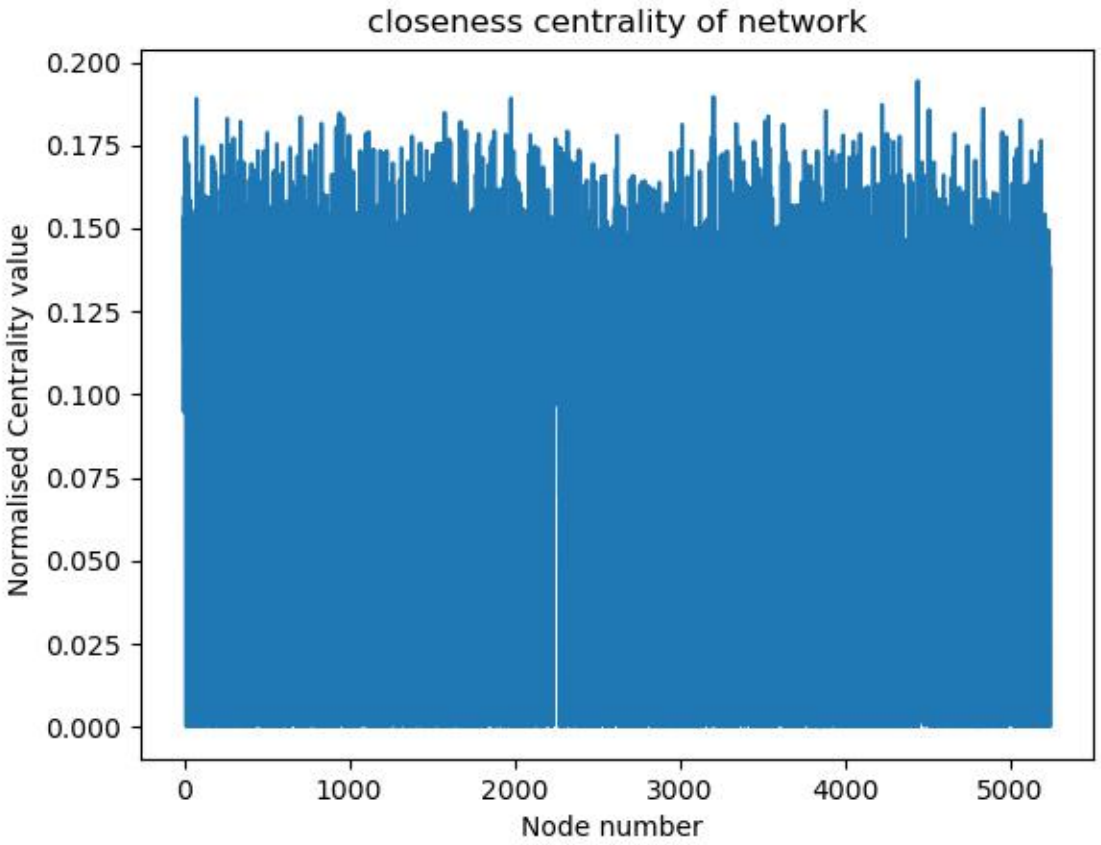
[2] General Relativity and Quantum Cosmology collaboration network

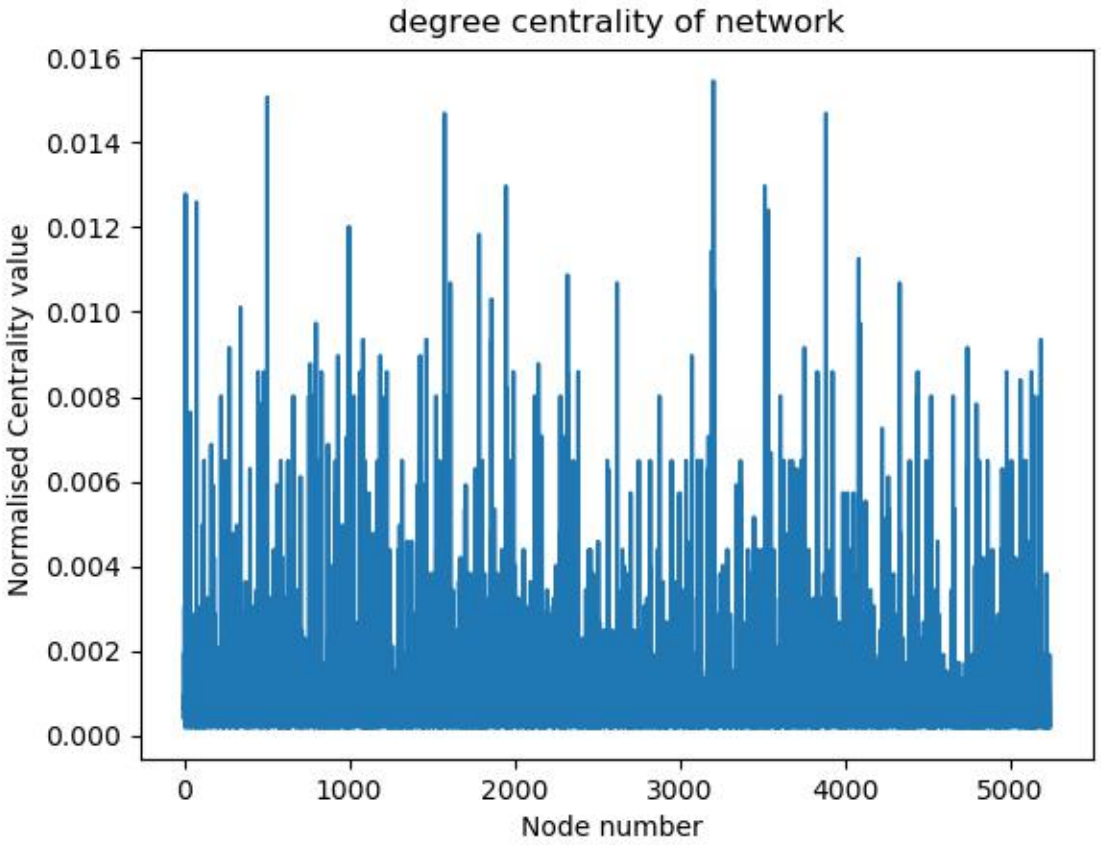
Dataset statistics	
Nodes	5242
Edges	14496
Nodes in largest WCC	4158 (0.793)
Edges in largest WCC	13428 (0.926)
Nodes in largest SCC	4158 (0.793)
Edges in largest SCC	13428 (0.926)
Average clustering coefficient	0.5296
Number of triangles	48260
Fraction of closed triangles	0.3619

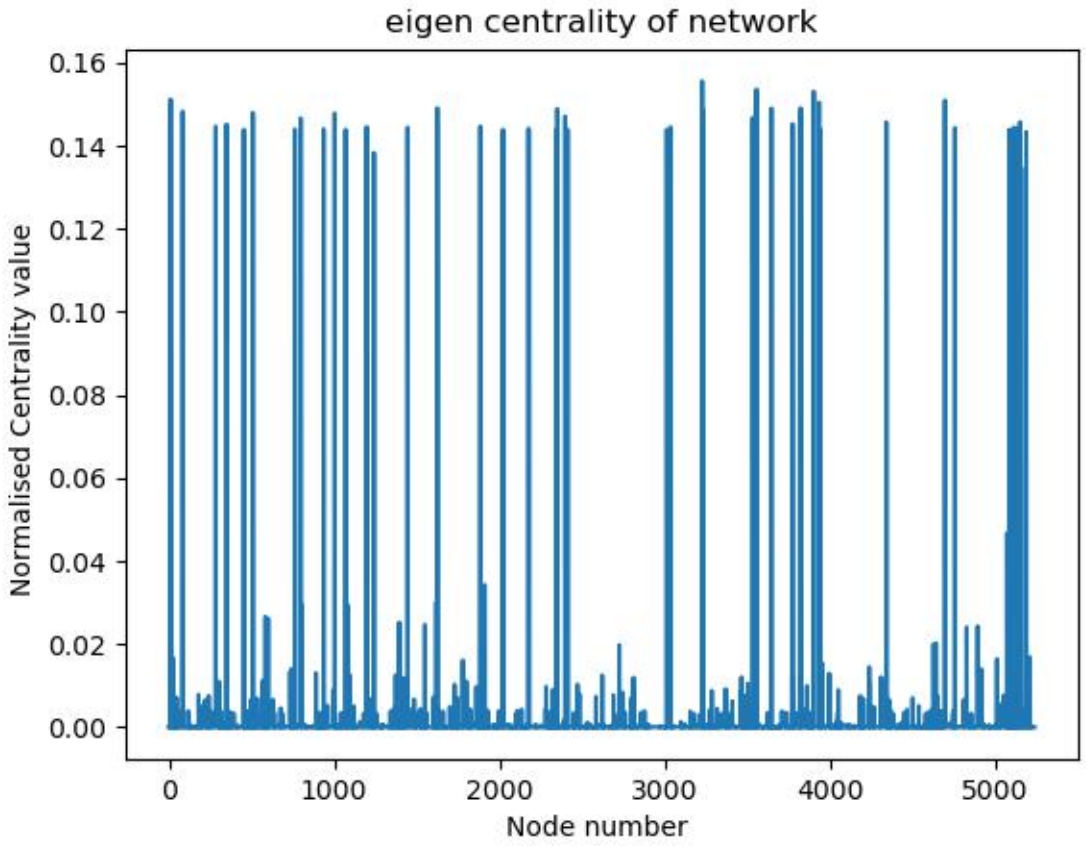
Diameter (longest shortest path)	17
90-percentile effective diameter	7.6

Below figures show the various centrality measures for this dataset




closeness centrality of network
Normalised Centrality value
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
1000
2000
3000
4000
5000
Node number


degree centrality of network
Normalised Centrality value
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
5000
Node number



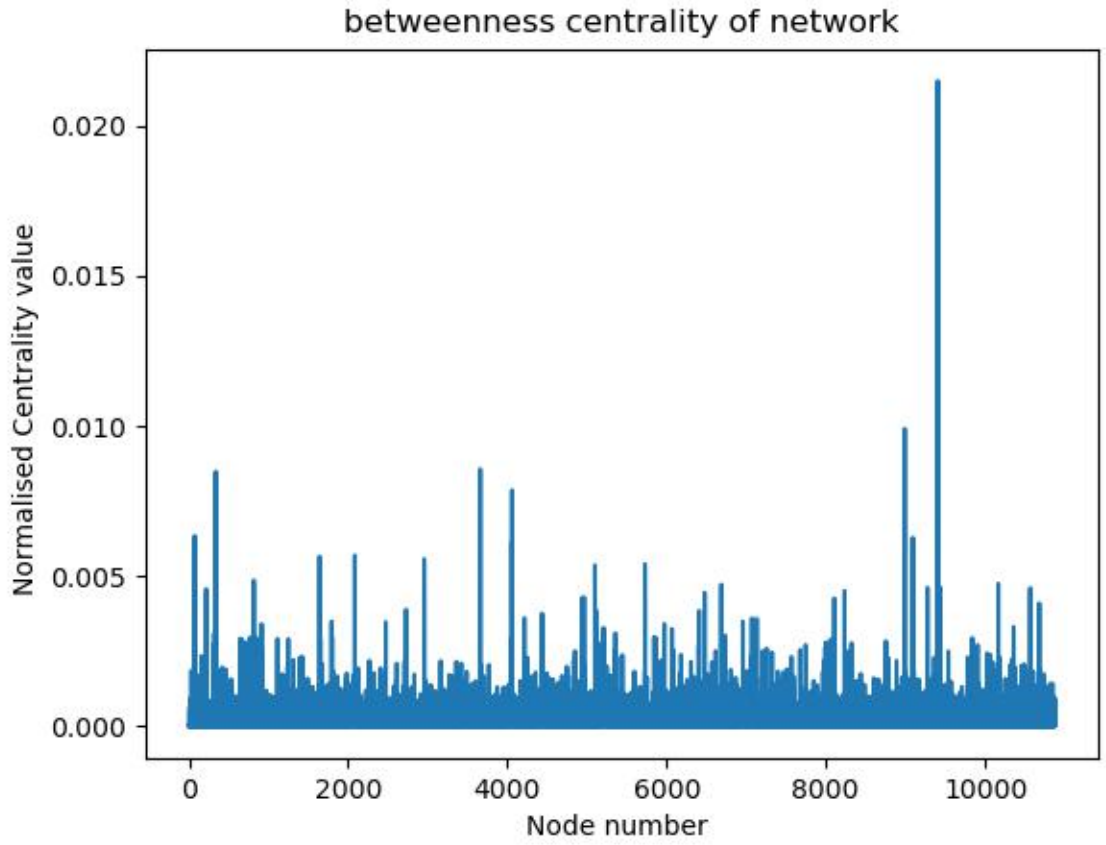
We can see that certain nodes have a very high eigen centrality measure compared to others. Even though other measures do not allow for us to see much difference. In this dataset eigenvector is the best centrality measure.

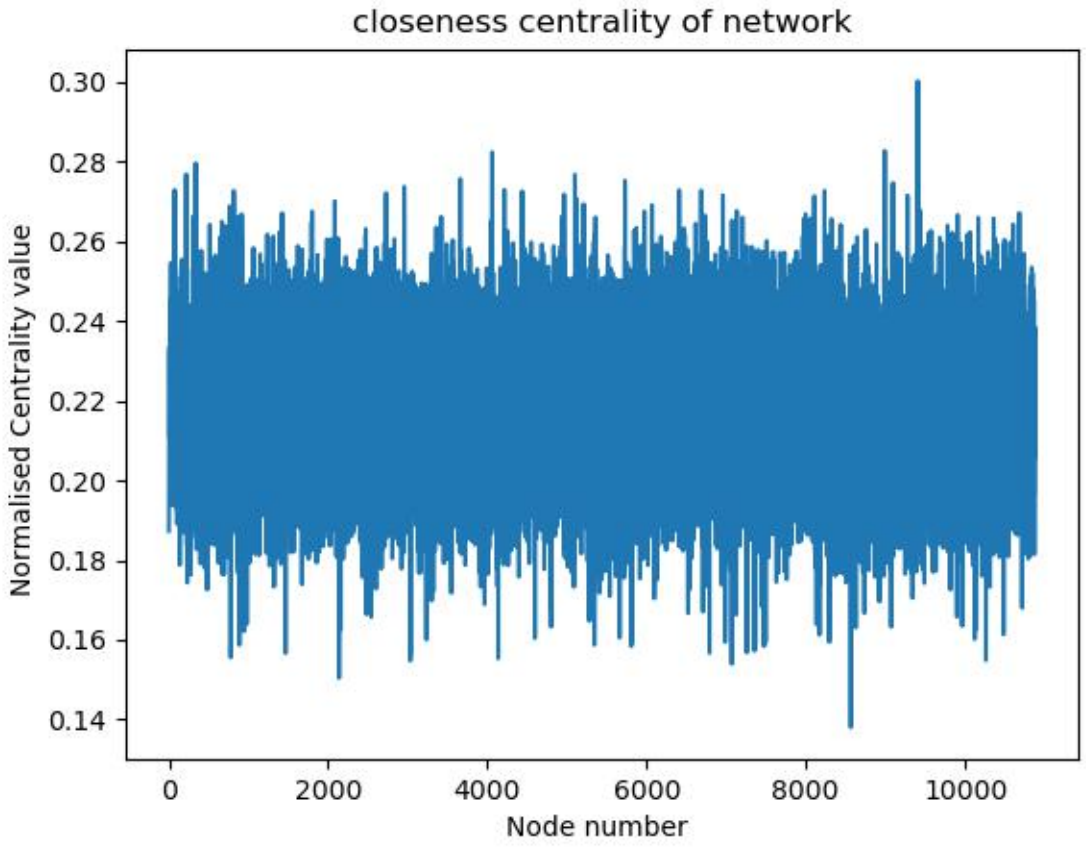
[3] Gnutella peer-to-peer network, August 4 2002

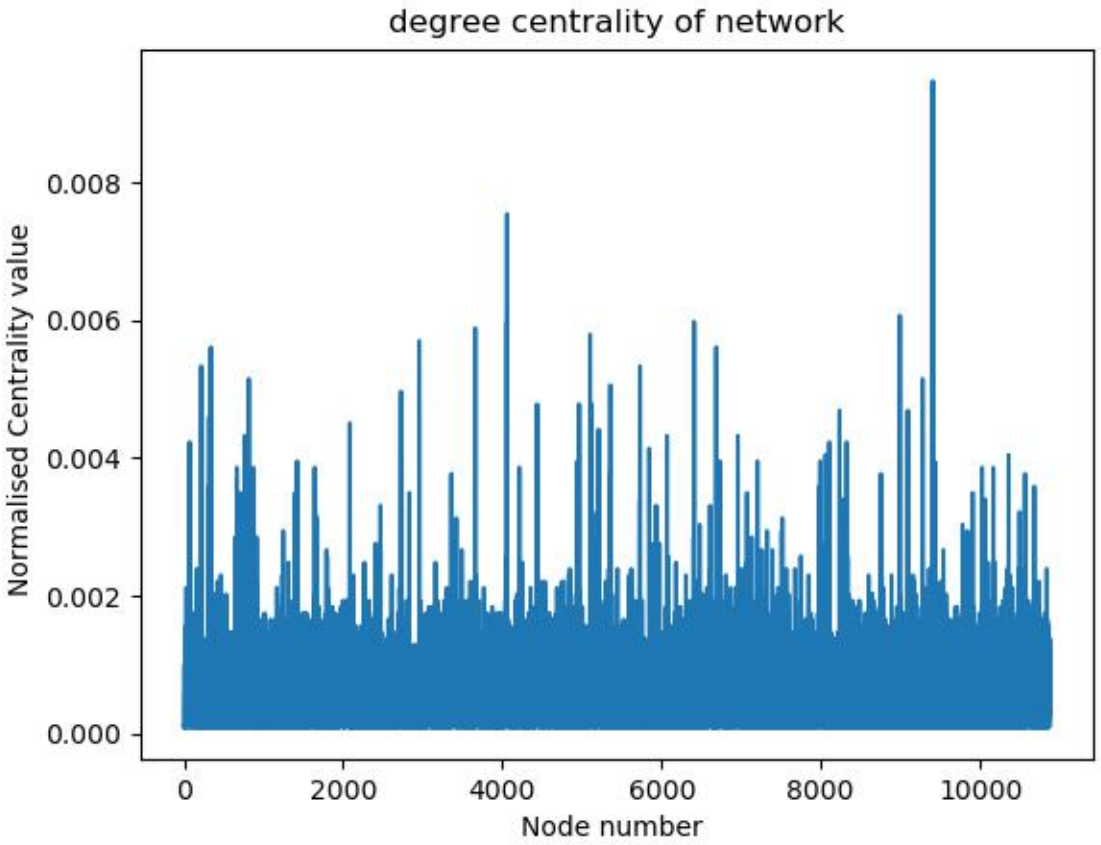
Dataset statistics	
Nodes	10876
Edges	39994
Nodes in largest WCC	10876 (1.000)
Edges in largest WCC	39994 (1.000)
Nodes in largest SCC	4317 (0.397)
Edges in largest SCC	18742 (0.469)
Average clustering coefficient	0.0062
Number of triangles	934
Fraction of closed triangles	0.001807
Diameter (longest shortest path)	9

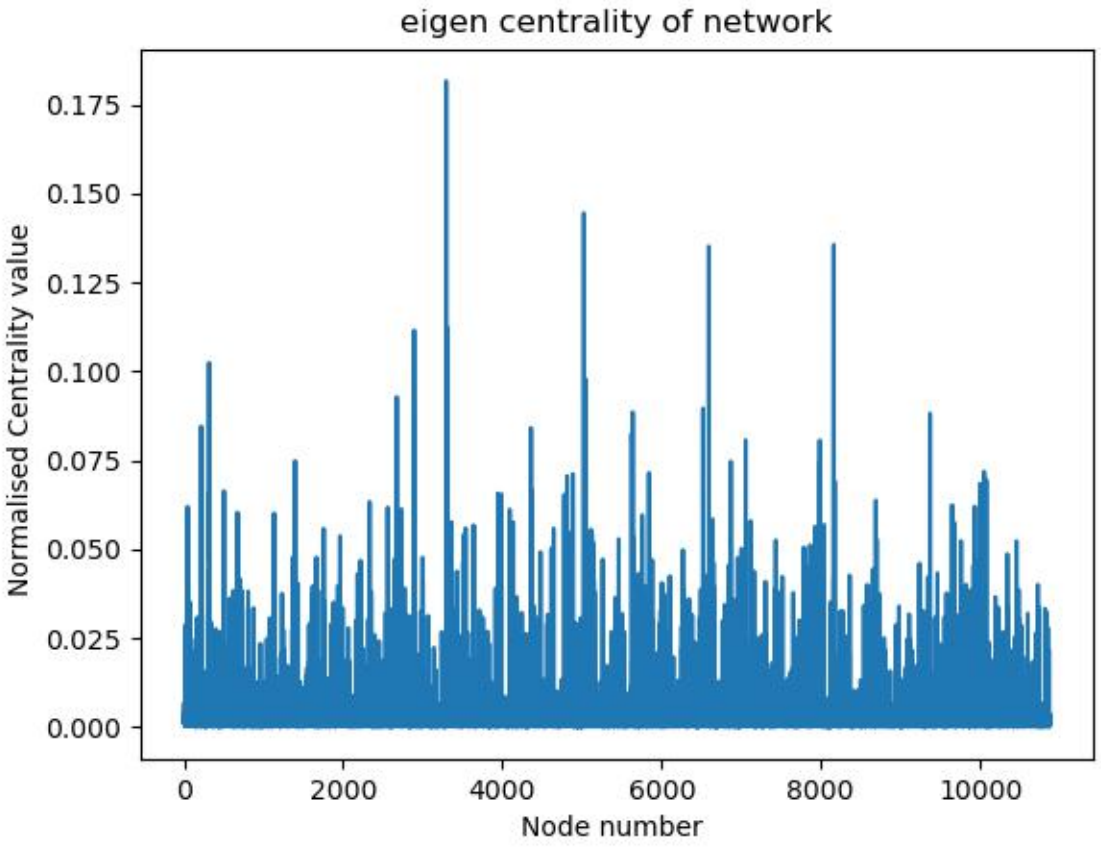
90-percentile effective diameter	5.4

Below are the plots of various centrality measures




closeness centrality of network
Normalised Centrality value
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0
2000
4000
6000
8000
10000
Node number


degree centrality of network
Normalised Centrality value
0.000
0.002
0.004
0.006
0.008
0
2000
4000
6000
8000
10000
Node number

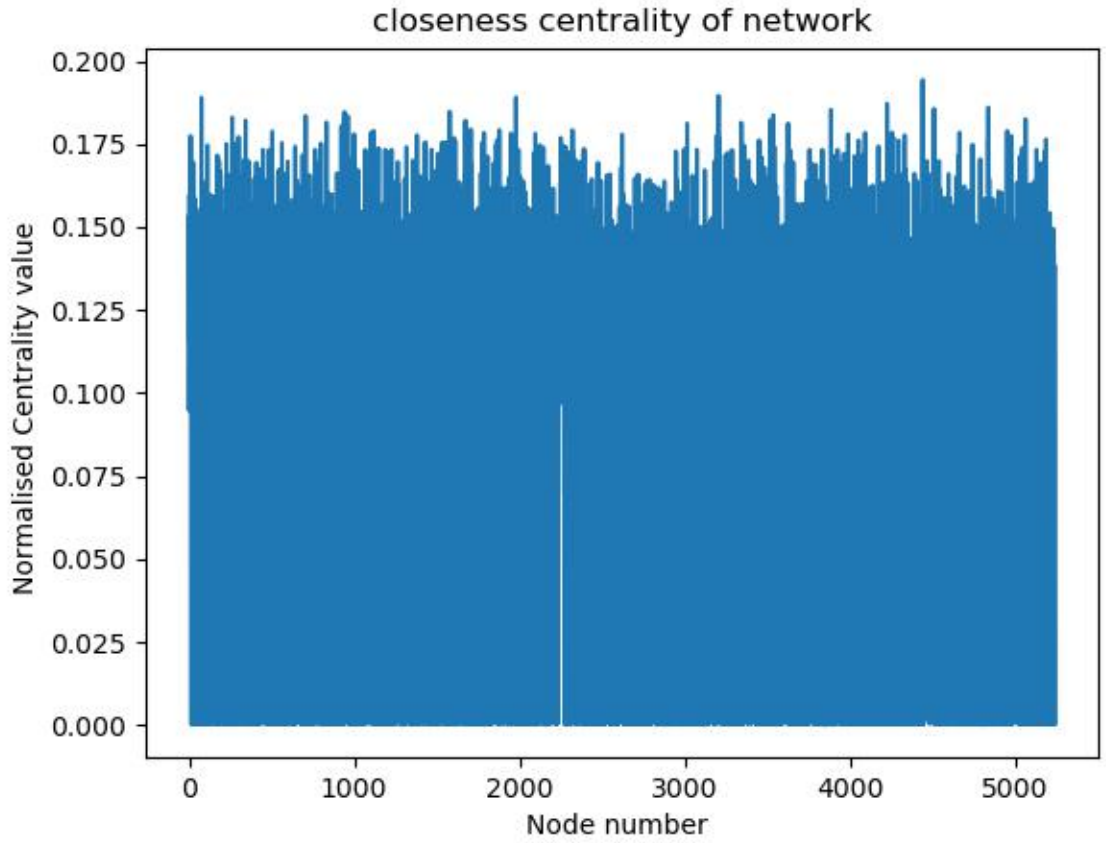


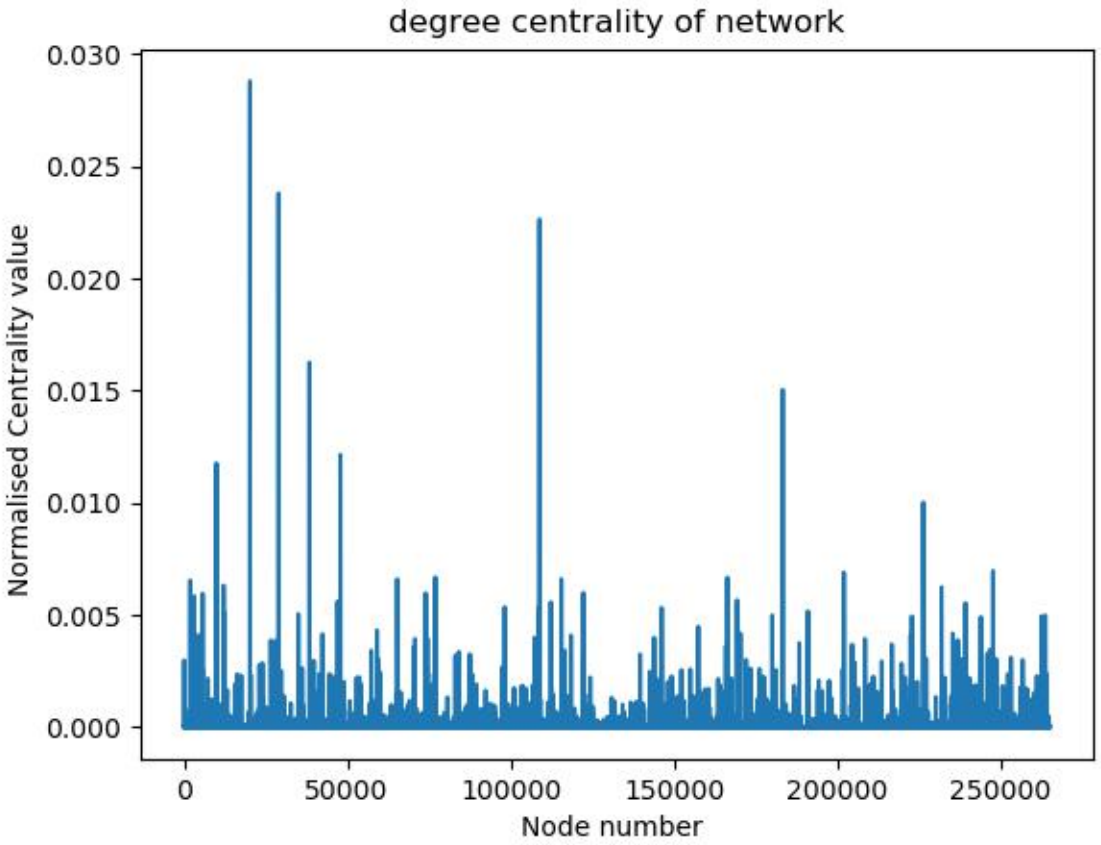
In this dataset betweenness centrality measure shows one node to be more important than any other by a large margin.

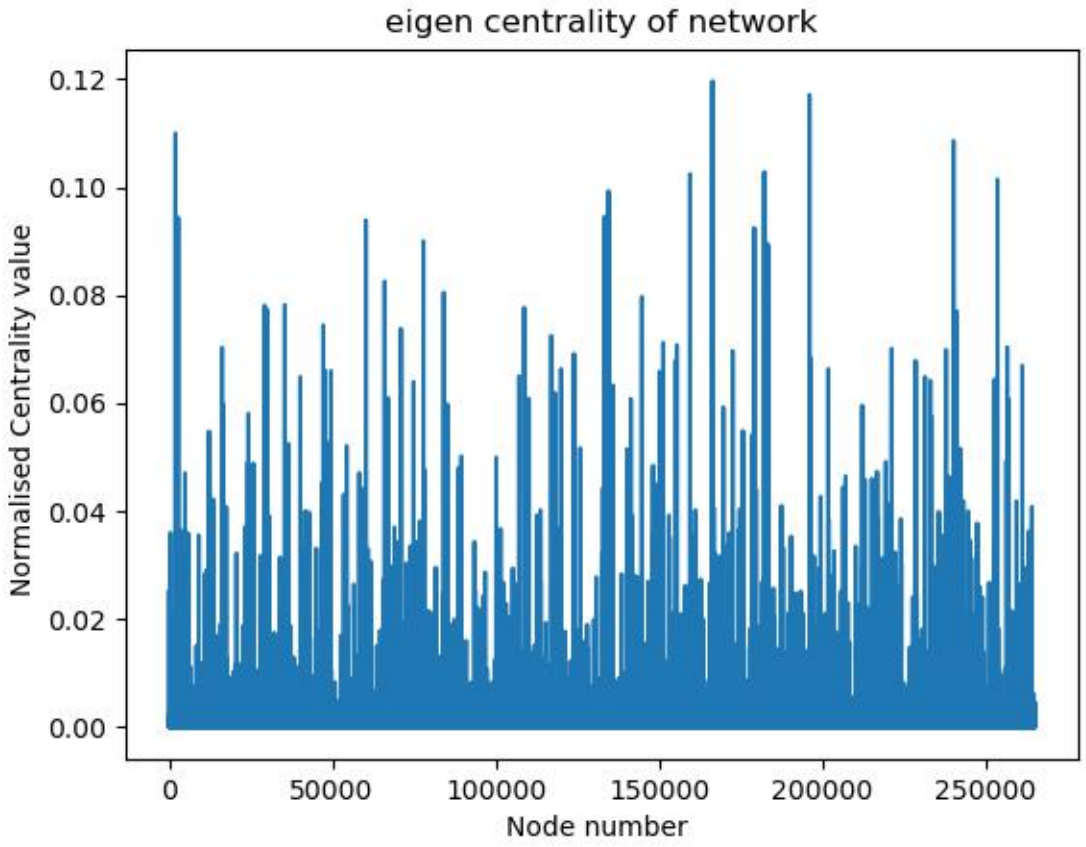
[4] EU email communication network

Dataset statistics	
Nodes	265214
Edges	420045
Nodes in largest WCC	224832 (0.848)
Edges in largest WCC	395270 (0.941)
Nodes in largest SCC	34203 (0.129)
Edges in largest SCC	151930 (0.362)
Average clustering coefficient	0.0671
Number of triangles	267313
Fraction of closed triangles	0.001373
Diameter (longest shortest path)	14
90-percentile effective diameter	4.5

Below are the plots for various centrality measures



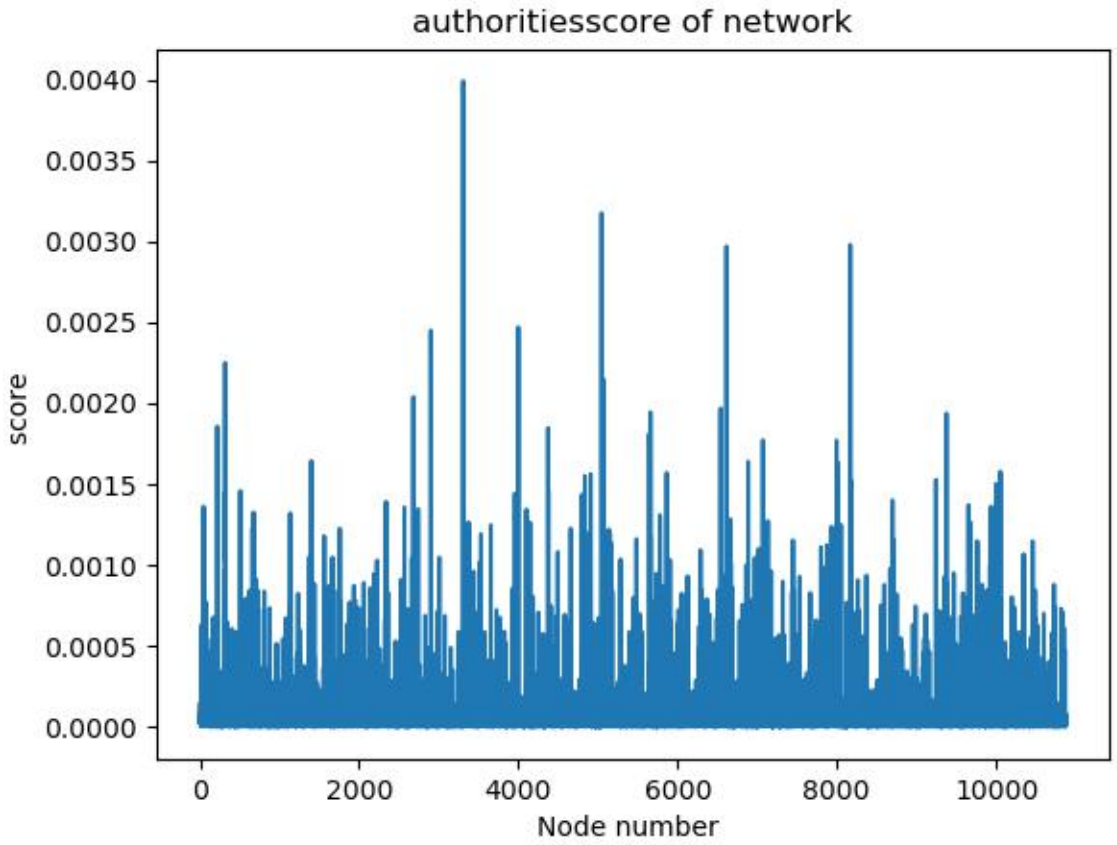

degree centrality of network
Normalised Centrality value
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50000
100000
150000
200000
250000
Node number

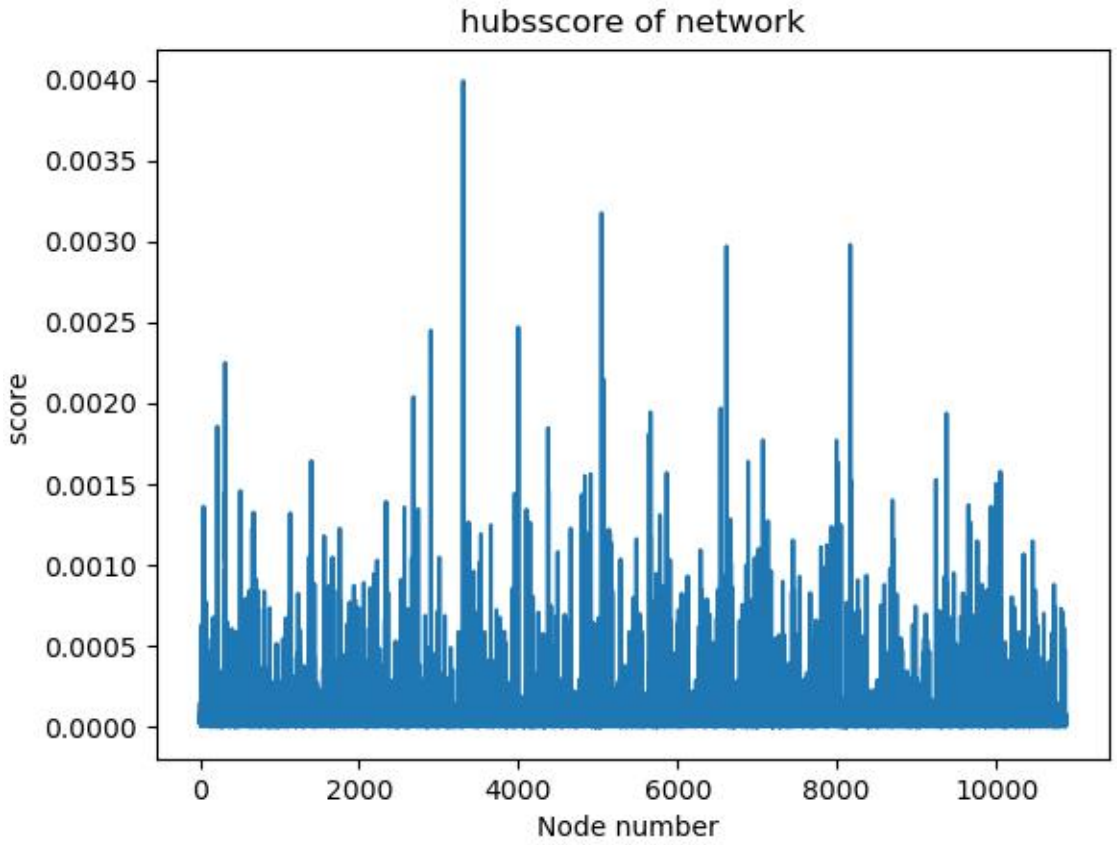


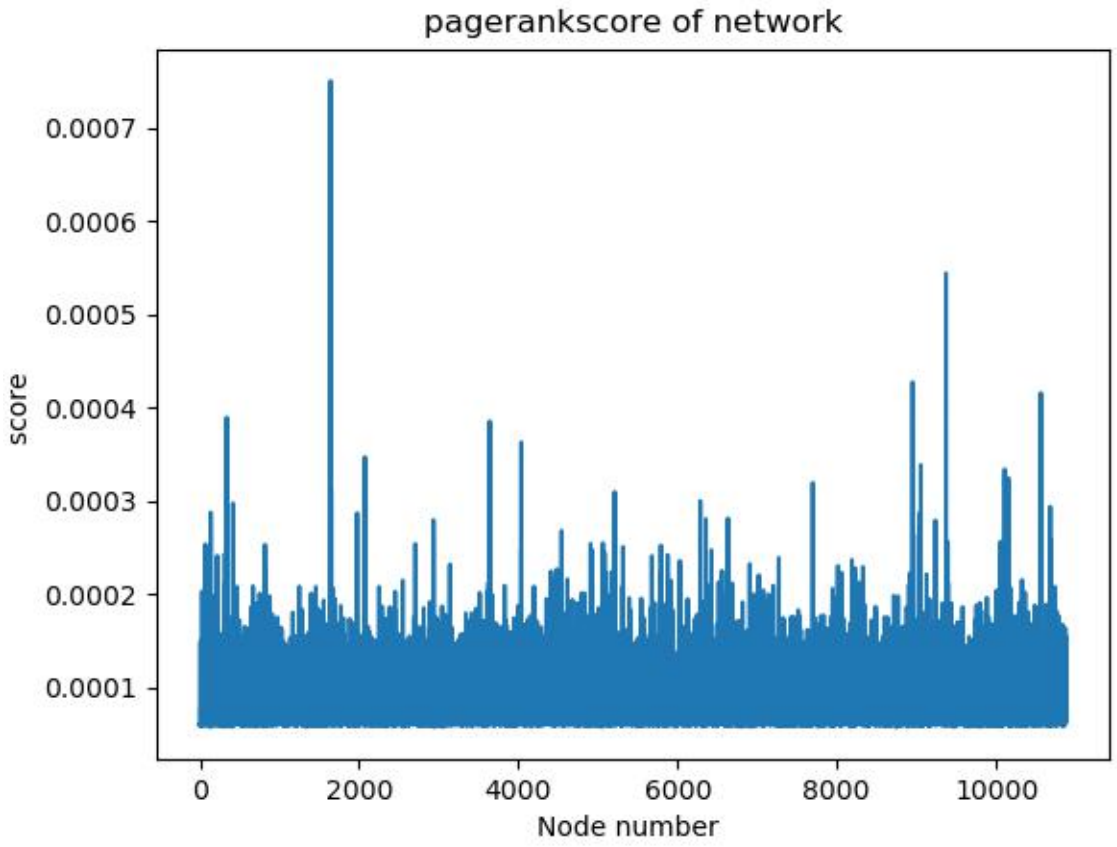
In this dataset we can see that there are are around 7 nodes with high number of degree and many more nodes with significantly higher eigen vector scores.

PAGERANK AND HITS

Page rank and hits was applied to the nutella dataset and below are the results


authoritiesscore of network
score
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
2000
4000
6000
8000
10000
Node number


hubsscore of network
score
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
2000
4000
6000
8000
10000
Node number



We can see that page rank has major outliers in the score whereas HITS is more evenly distributed.

The HITS score seems to be similar to the eigen score we saw in the previous experiments.

The pagerank seems to giving more importance node in between 1500-2000 whereas none of the centrality metrics or HITS shows the same.

HIGHEST SCORE NODE

[1] HITS: Node - 1054 value -0.0039

[2] PageRank: Node - 5598 value - 0.00075

LOWEST SCORE NODE

[1] HITS: Node - 10215 value -1.23697084e-08

[2] PageRank: Node- 2847 value- 5.6946960e-08

CONCLUSION

In this we explored various methods of measuring importance of nodes in social networks and ranking web pages. We experimented with various size datasets to see their properties and analyse them.

References

[1] https://www.geeksforgeeks.org/eigenvector-centrality-centrality-measure/

[2] https://www.geeksforgeeks.org/betweenness-centrality-centrality-measure/

[3] https://en.wikipedia.org/wiki/PageRank

[4] https://en.wikipedia.org/wiki/HITS_algorithm

[5] https://snap.stanford.edu/data/