# ASSIGNMENT 3

*Giridhar S(MT21026)* *Alka Kumari(MT21006)*

## QUESTION1:

## Dataset:

The task is performed on the "`soc-sign-bitcoinalpha.csv.gz`" dataset. The CSV file contains 4 columns. The first column contains the source node, the second column contains the target node, the third column contains the ratings and the fourth column contains the timestamps.

## Pre-processing Steps:

This is the first step that cleans the data in the dataset. This process includes operations to remove unnecessary data, handle missing data, normalize the data, and also to remove redundancy in the dataset.

1. Firstly, all the useful libraries were imported.

2. Then, the dataset was loaded in the form of a data frame.

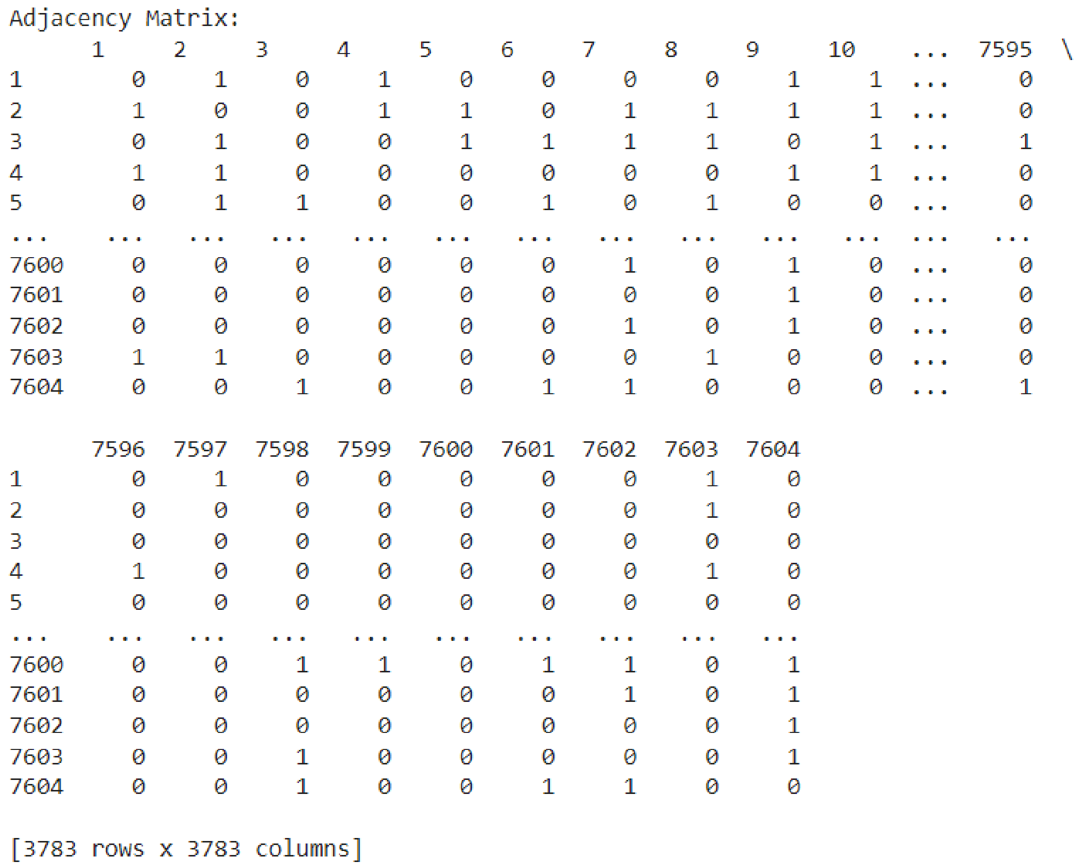
3. Then, the columns like rating and time were dropped.

4. All the duplicates were removed.

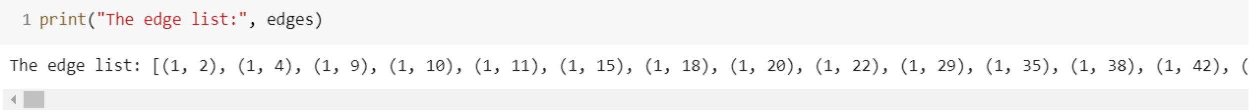
## Methodology:

1. First, all the unique nodes and edges were found from the "Source" and "Target" columns and stored in nodes and edges respectively.

2. Next, using the nodes and edges, an adjacency matrix was constructed in the form of a data frame and the edge list was constructed as the list of edges.



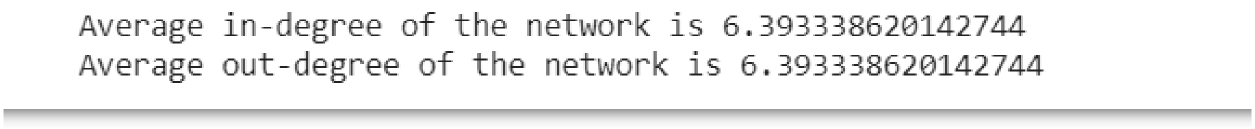
```
Adjacency Matrix:
      1     2     3     4     5     6     7     8     9     10    ...  7595  \
1        0     1     0     1     0     0     0     0     1     1  ...     0
2        1     0     0     1     1     0     1     1     1     1  ...     0
3        0     1     0     0     1     1     1     1     0     1  ...     1
4        1     1     0     0     0     0     0     0     1     1  ...     0
5        0     1     1     0     0     1     0     1     0     0  ...     0
...    ...   ...   ...   ...   ...   ...   ...   ...   ...   ...  ...   ...
7600     0     0     0     0     0     0     1     0     1     0  ...     0
7601     0     0     0     0     0     0     0     0     1     0  ...     0
7602     0     0     0     0     0     0     1     0     1     0  ...     0
7603     1     1     0     0     0     0     0     1     0     0  ...     0
7604     0     0     1     0     0     1     1     0     0     0  ...     1

      7596  7597  7598  7599  7600  7601  7602  7603  7604
1        0     1     0     0     0     0     0     1     0
2        0     0     0     0     0     0     0     1     0
3        0     0     0     0     0     0     0     0     0
4        1     0     0     0     0     0     0     1     0
5        0     0     0     0     0     0     0     0     0
...    ...   ...   ...   ...   ...   ...   ...   ...   ...
7600     0     0     1     1     0     1     1     0     1
7601     0     0     0     0     0     0     1     0     1
7602     0     0     0     0     0     0     0     0     1
7603     0     0     1     0     0     0     0     0     1
7604     0     0     1     0     0     1     1     0     0

[3783 rows x 3783 columns]
```



```
print("The edge list:", edges)
```

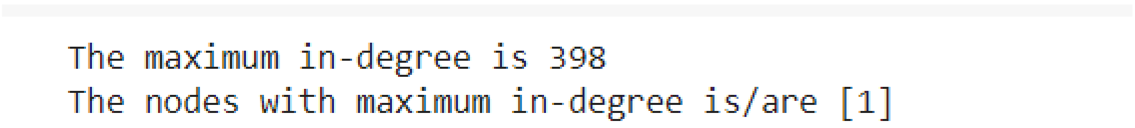
```
The edge list: [(1, 2), (1, 4), (1, 9), (1, 10), (1, 11), (1, 15), (1, 18), (1, 20), (1, 22), (1, 29), (1, 35), (1, 38), (1, 42), (
```

3. Two dictionaries were made namely in_deg and out_deg, one to store indegree and other to store outdegree of each node. From this, average indegree and outdegree were calculated which turned out to be equal.

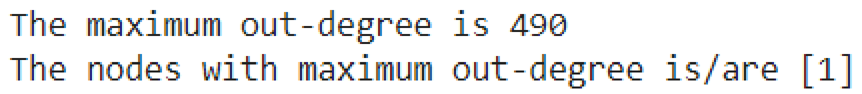


```
Average in-degree of the network is 6.393338620142744
Average out-degree of the network is 6.393338620142744
```

4. The above two dictionaries were sorted in descending order to fetch the nodes with maximum indegree and outdegree respectively.

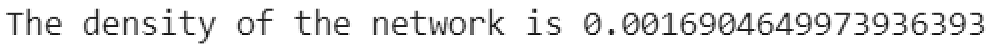


```
The maximum in-degree is 398
The nodes with maximum in-degree is/are [1]
```



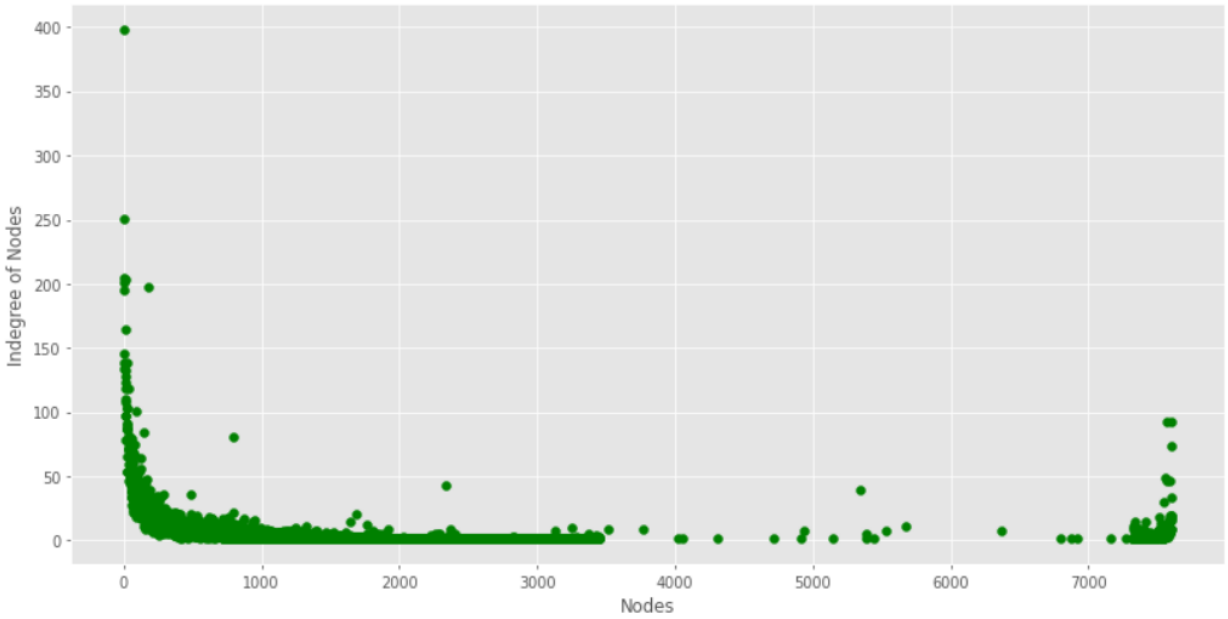
```
The maximum out-degree is 490
The nodes with maximum out-degree is/are [1]
```

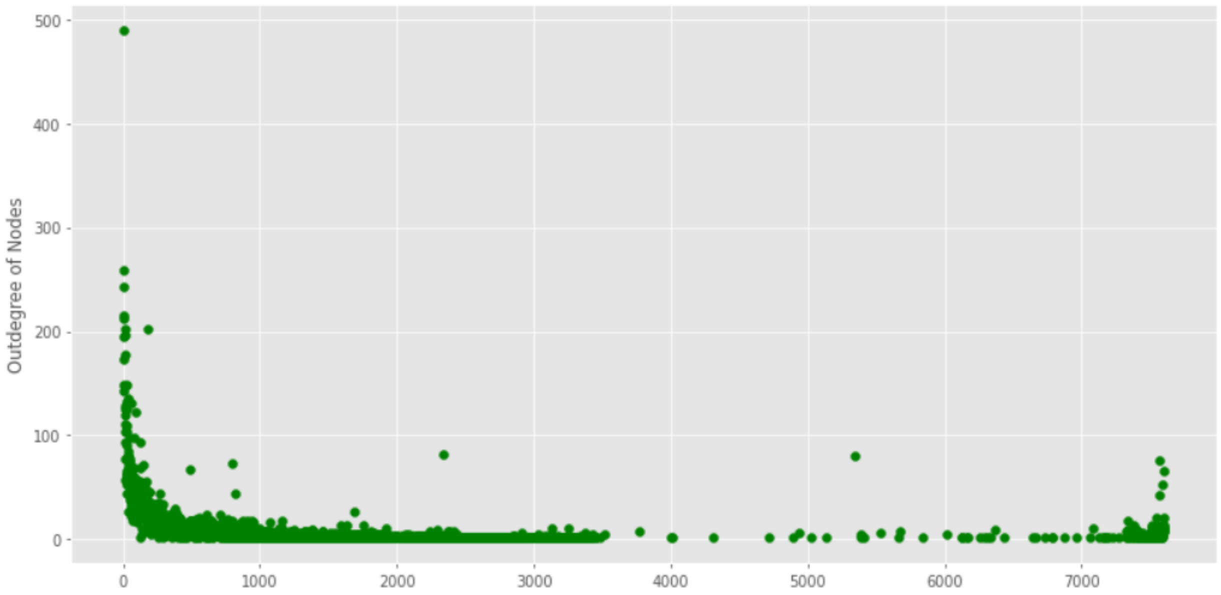
5. Next, the density of the network was calculated.

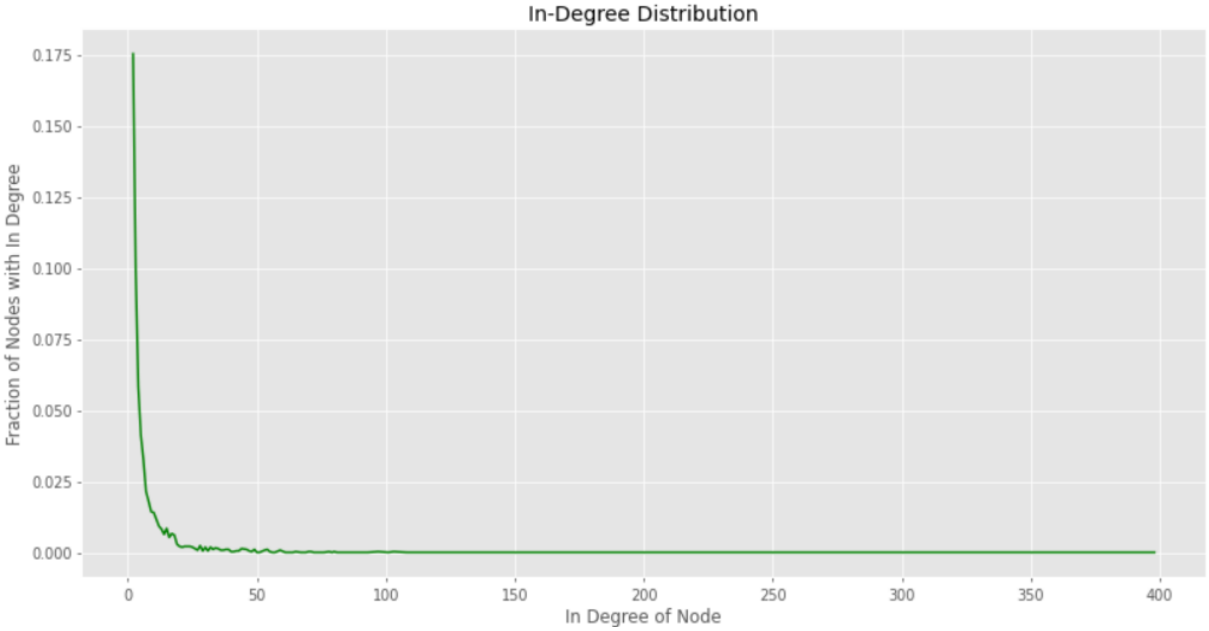


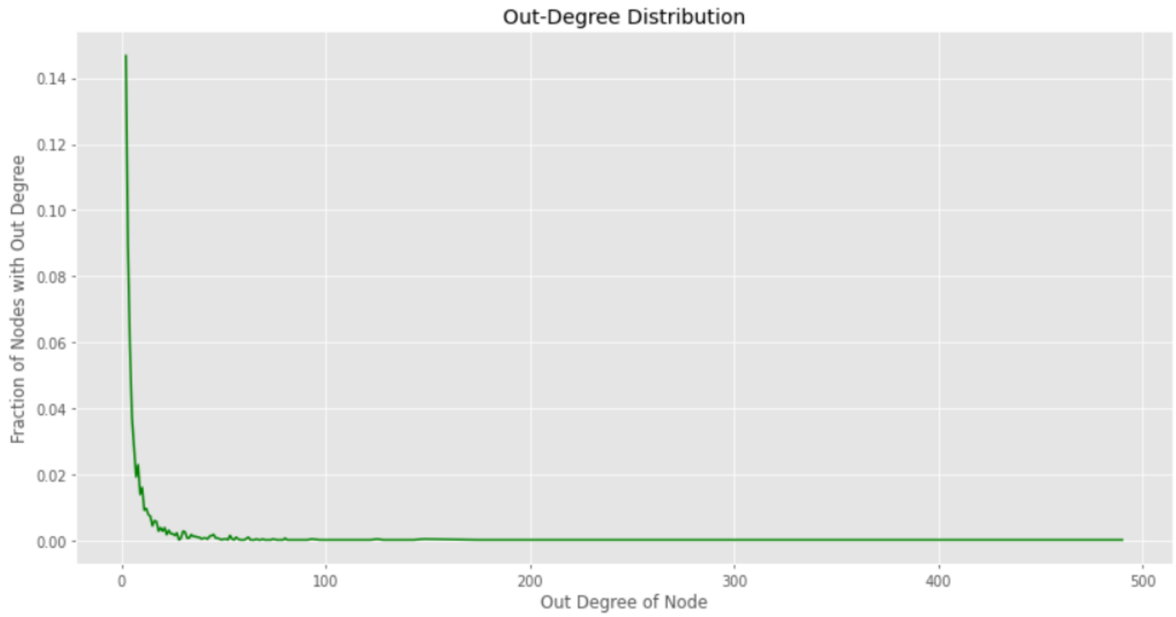
```
The density of the network is 0.0016904649973936393
```

6. Next, the indegree and outdegree distribution was plotted.

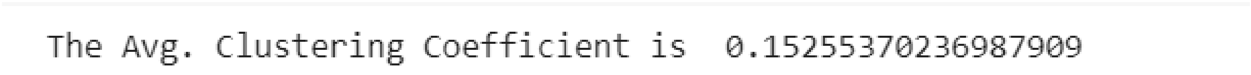



Outdegree of Nodes
500
400
300
200
100
0
0
1000
2000
3000
4000
5000
6000
7000


In-Degree Distribution
Fraction of Nodes with In Degree
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
50
100
150
200
250
300
350
400
In Degree of Node

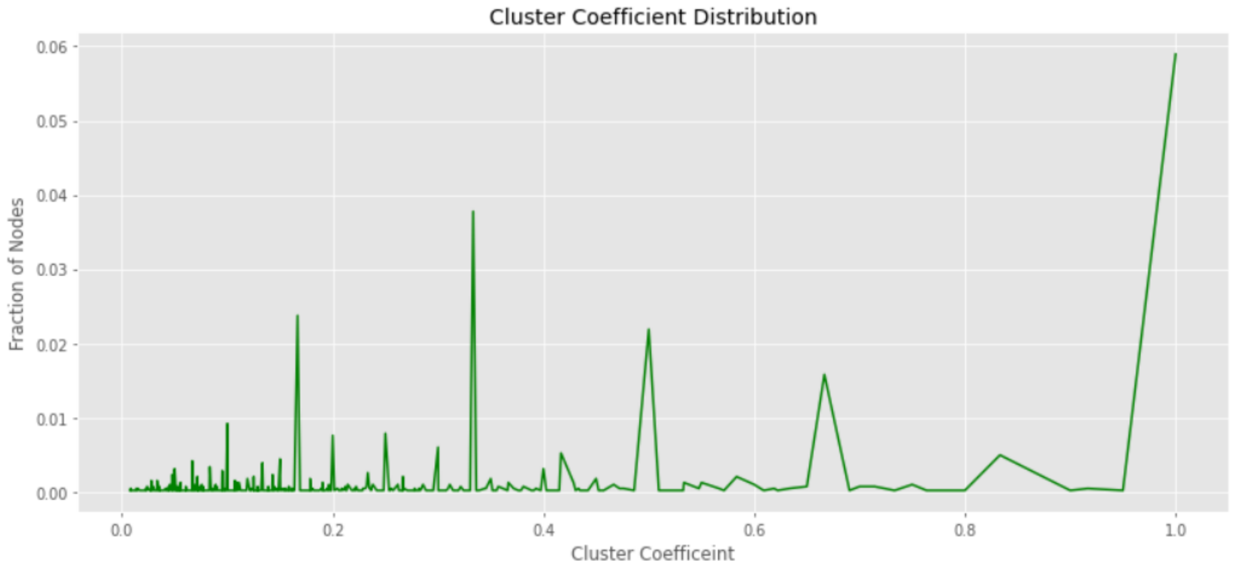


7. Then, the clustering coefficient for each node was calculated and then the average clustering coefficient was calculated.



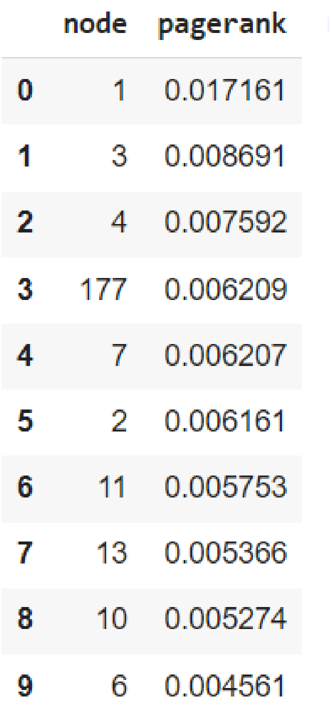
```
The Avg. Clustering Coefficient is  0.15255370236987909
```

8. Then, the clustering coefficient distribution was plotted.



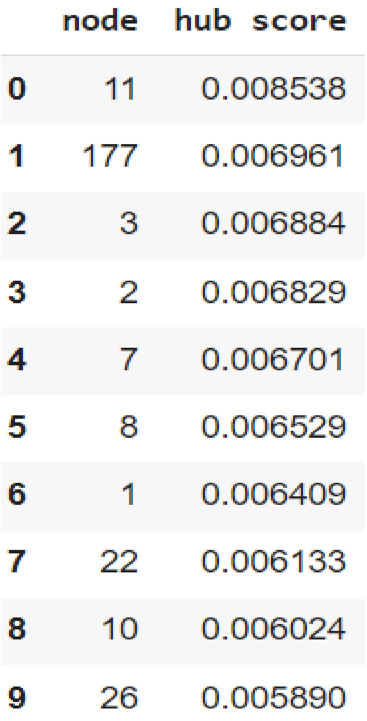
# QUESTION2:

1. The pagerank score for each node is calculated using the networkx library. The top 10 nodes with high pagerank scores are shown below.



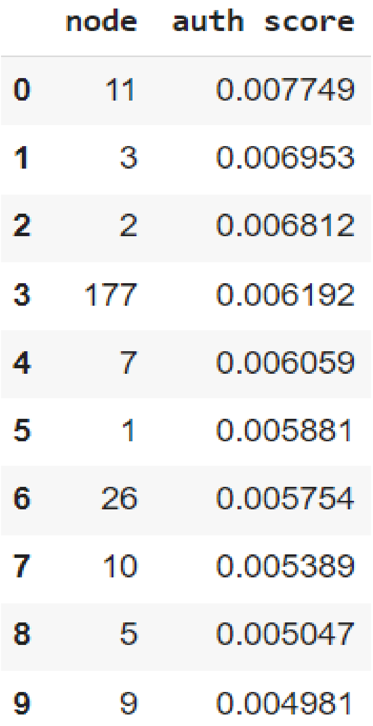
| | node | pagerank |
|---|---|---|
| 0 | 1 | 0.017161 |
| 1 | 3 | 0.008691 |
| 2 | 4 | 0.007592 |
| 3 | 177 | 0.006209 |
| 4 | 7 | 0.006207 |
| 5 | 2 | 0.006161 |
| 6 | 11 | 0.005753 |
| 7 | 13 | 0.005366 |
| 8 | 10 | 0.005274 |
| 9 | 6 | 0.004561 |

2. The hub score for each node is calculated using networkx and the top 10 nodes with high hub score are shown below.



| | node | hub score |
|---|---|---|
| 0 | 11 | 0.008538 |
| 1 | 177 | 0.006961 |
| 2 | 3 | 0.006884 |
| 3 | 2 | 0.006829 |
| 4 | 7 | 0.006701 |
| 5 | 8 | 0.006529 |
| 6 | 1 | 0.006409 |
| 7 | 22 | 0.006133 |
| 8 | 10 | 0.006024 |
| 9 | 26 | 0.005890 |

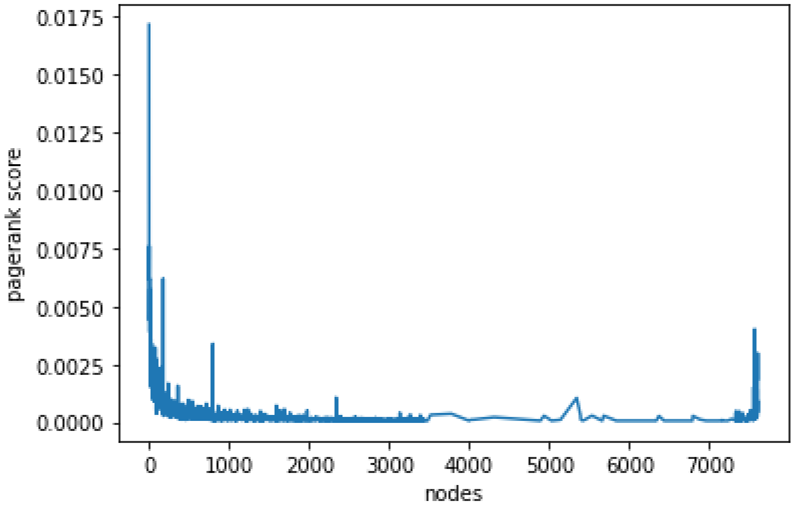
The auth score for each node is calculated using networkx and the top 10 nodes with high auth score are shown below.



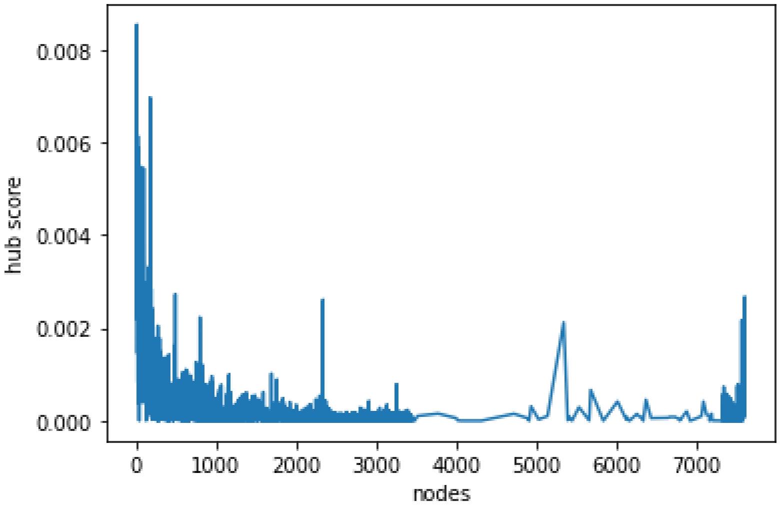
| | node | auth score |
|---|---|---|
| 0 | 11 | 0.007749 |
| 1 | 3 | 0.006953 |
| 2 | 2 | 0.006812 |
| 3 | 177 | 0.006192 |
| 4 | 7 | 0.006059 |
| 5 | 1 | 0.005881 |
| 6 | 26 | 0.005754 |
| 7 | 10 | 0.005389 |
| 8 | 5 | 0.005047 |
| 9 | 9 | 0.004981 |

The top 10 pagerank scores and the hub scores can be compared. In case of pagerank score, node 1 is the most influential node whereas in case of hub score, node 11 is the node with the highest hub score, i.e. it has points to many authority nodes.

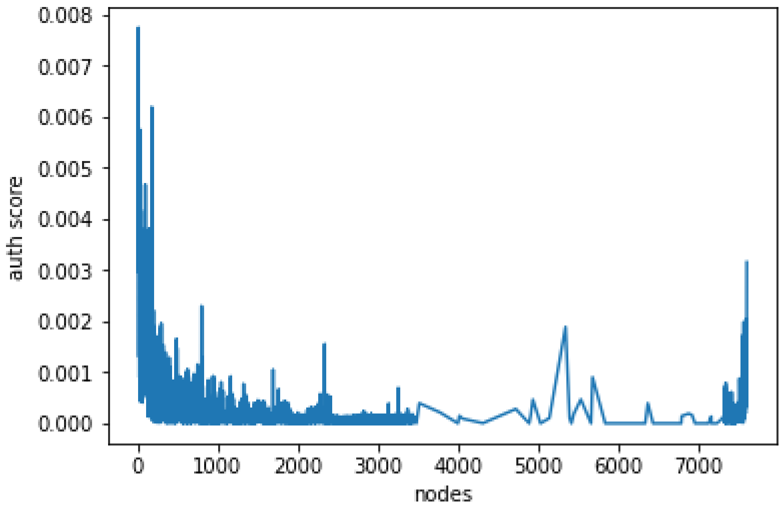
The distribution of pagerank score is shown below:



The distribution of hub score is shown below:



The distribution of auth score is given below:



It can be seen that the distribution is more or less the same in all the three scores.