Assignment 3 Report

We have selected the **p2p-Gnutella08.txt** dataset for our work with 6301 nodes and 20777 edges. The dataset corresponds to a directed graph.

Question 1:

a). Graph representation as an adjacency list

```
{0: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
 3: [703, 826, 1097, 1287, 1591, 1895, 1896, 1897, 1898, 1899],
 4: [144, 258, 491, 1021, 1418, 1669, 1900, 1901, 1902, 1903],
 5: [121, 127, 128, 179, 247, 249, 264, 353, 424, 426],
 7: [145, 176, 177, 353, 753, 754, 762, 2064, 3002],
 8: [520, 665, 852, 1394, 1786, 1842, 1904, 1905, 1906, 1907],
 9: [124, 147, 177, 246, 247, 248, 249, 250, 251, 252],
 11: [12, 13, 14, 15, 16, 17, 18, 19, 20, 21],
```

b). Graph representation as adjacency matrix:

```
adjacency matrix
array([[0, 1, 1, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

c). No. of vertices and edges in the graph

```
print ("no. of edges:", len(edges))
print ("no. of vertices:", len(vertices))
```

```
no. of edges: 20777
no. of vertices: 6301
```

d). Average out-degree and in-degree of graph

We calculated the total in-degree and out-degree and divided both the values with number of nodes in the graph.

```
Total outdegree : 20777
Total vertices : 6301
Average outdegree : 3.2974131090303125
```

```
Total indegree : 20777
Total vertices : 6301
Average indegree : 3.2974131090303125
```

e). Nodes with max in-degree and out-degree along with values.

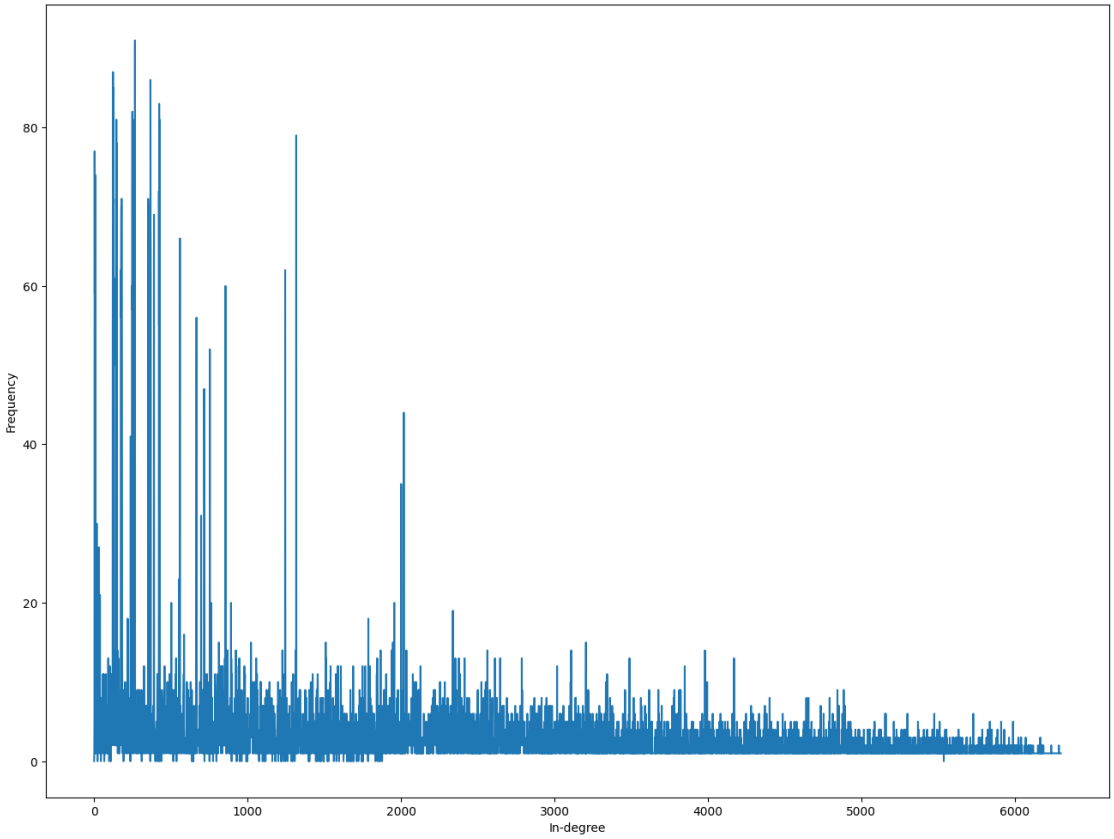
```
maximum In-degree : 91
vertex with Maximum In-degree : 266
Maximum Out-Degree : 48
vertex with Maximum Out-degree  5831
```

f). Network density is the ratio of (edges/maximum possible edges).

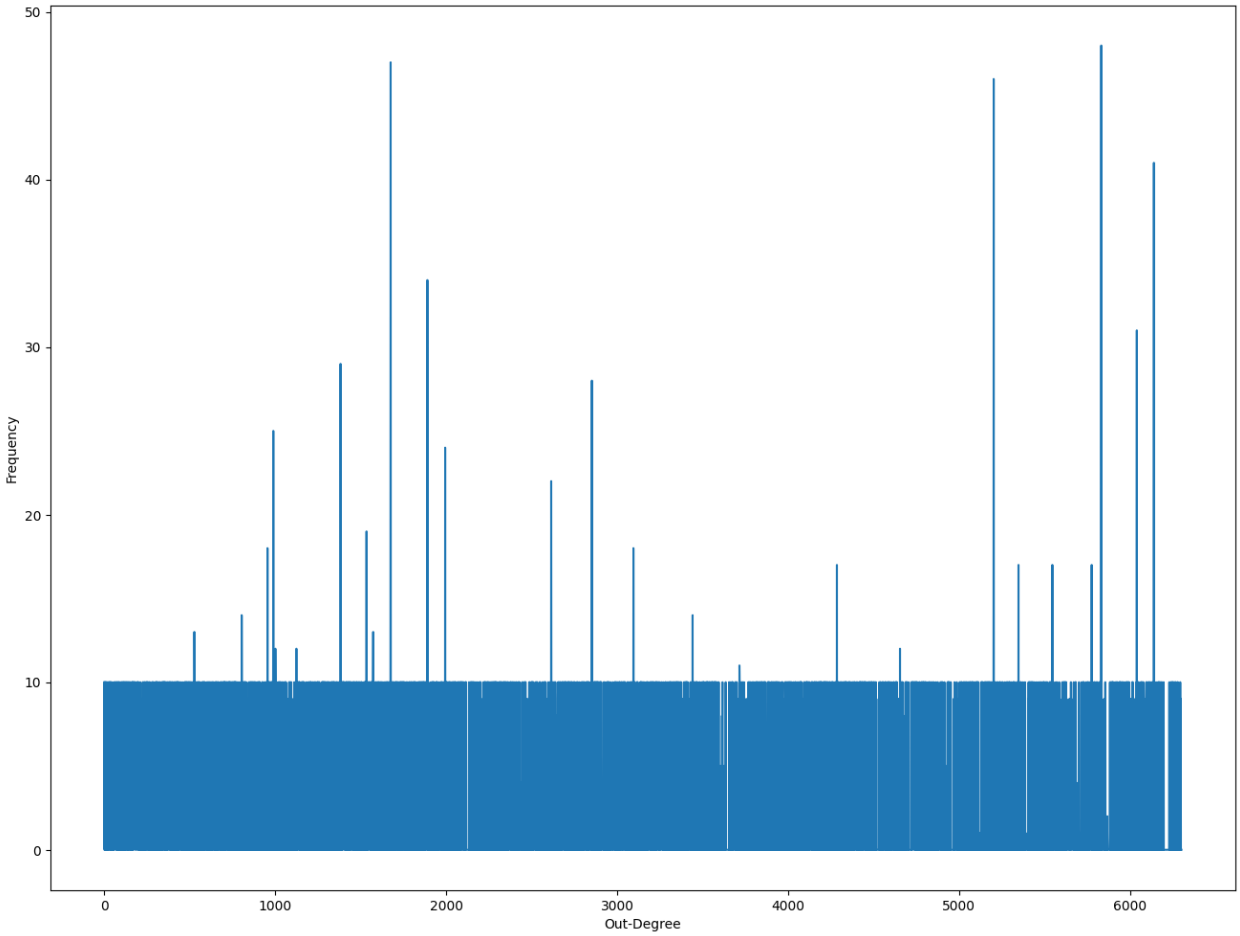
```
maximum edges possible for complete graph:  39696300
total edges in graph  20777
Network Density is 0.0005233989061952878
```

g).

Plot in-degree and out-degree distribution

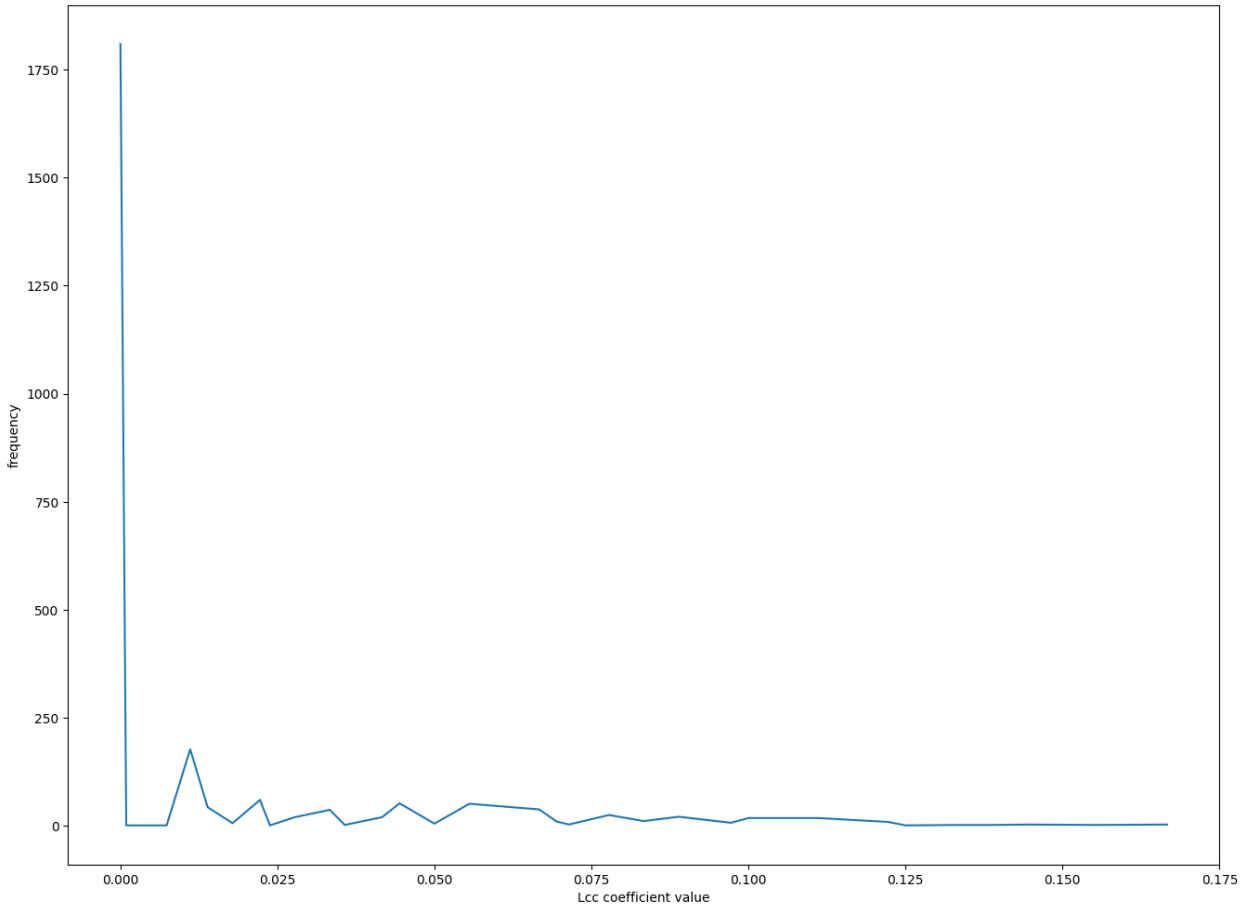


Plot out-degree



h). Calculate the local clustering coefficient of each node and plot the clustering-coefficient distribution (lcc vs frequency of lcc) of the network.

To calculate lcc, we have used the adjacency list to find the immediate neighbour of the given node and calculated the ratio of total links between the neighbours divided by the maximum link possible.



Question 2:

Methodology:

- Created an empty object of networkx directed graph.
- Parsed the dataset line by line and added each edge to the graph.
- Used the pagerank() and hits() function on the graph.
- Created a dataframe to report the pagerank, authority and hub score of each node.
- Highest PageRank score is of Node 367 and it is 0.00237.
- Highest Authority score is of Node 367 and it is 0.0214.
- Highest PagHubeRank score is of Node 3459 and it is 0.003032.

Reporting the score of first 20 Nodes:

	Node	PageRank	Authority	Hub
0	0	0.00010	0.00000	0.00148
1	1	0.00011	0.00015	0.00000
2	2	0.00017	0.00075	0.00000
3	3	0.00142	0.01859	0.00000
4	4	0.00154	0.01248	0.00029
5	5	0.00182	0.01648	0.00261
6	6	0.00011	0.00015	0.00000
7	7	0.00150	0.01135	0.00138
8	8	0.00126	0.01542	0.00001
9	9	0.00120	0.01059	0.00233
10	10	0.00031	0.00126	0.00000
11	11	0.00014	0.00000	0.00009
12	12	0.00011	0.00001	0.00000
13	13	0.00017	0.00001	0.00000
14	14	0.00016	0.00001	0.00000
15	15	0.00018	0.00001	0.00271
16	16	0.00012	0.00001	0.00000
17	17	0.00062	0.00494	0.00000
18	18	0.00015	0.00001	0.00000
19	19	0.00011	0.00001	0.00000

Comparison of PageRank and HITS method:

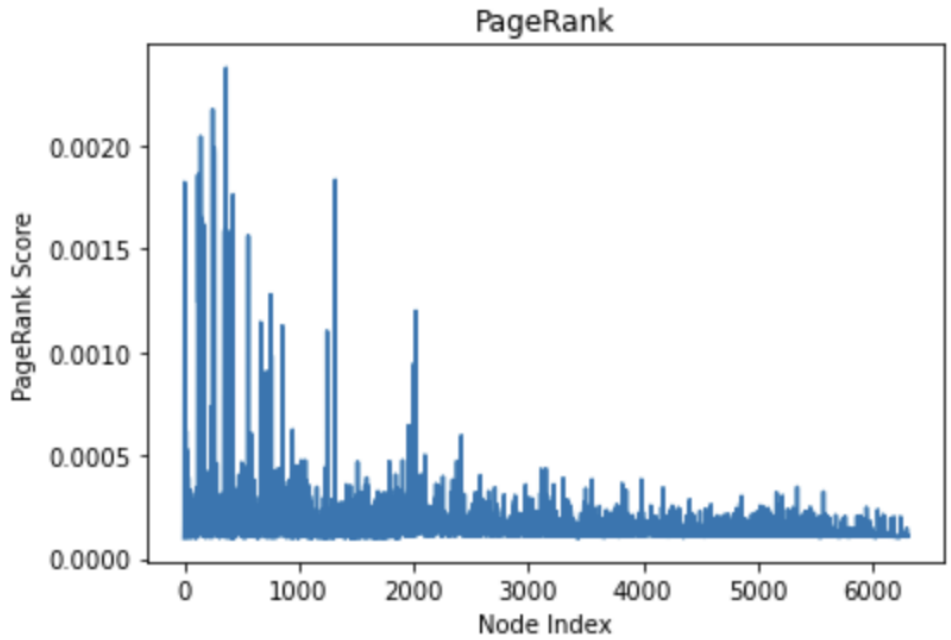
A node with high pagerank score indicates that the node has many quality links directing towards it. And during a random walk there is a high probability of spending time at the particular node.

A node with high authority score shows that there are many nodes pointing/directing towards it (high number of incoming links). Not only these nodes are high in number but they also have high hub score. This boosts the authority of the node in comparison.

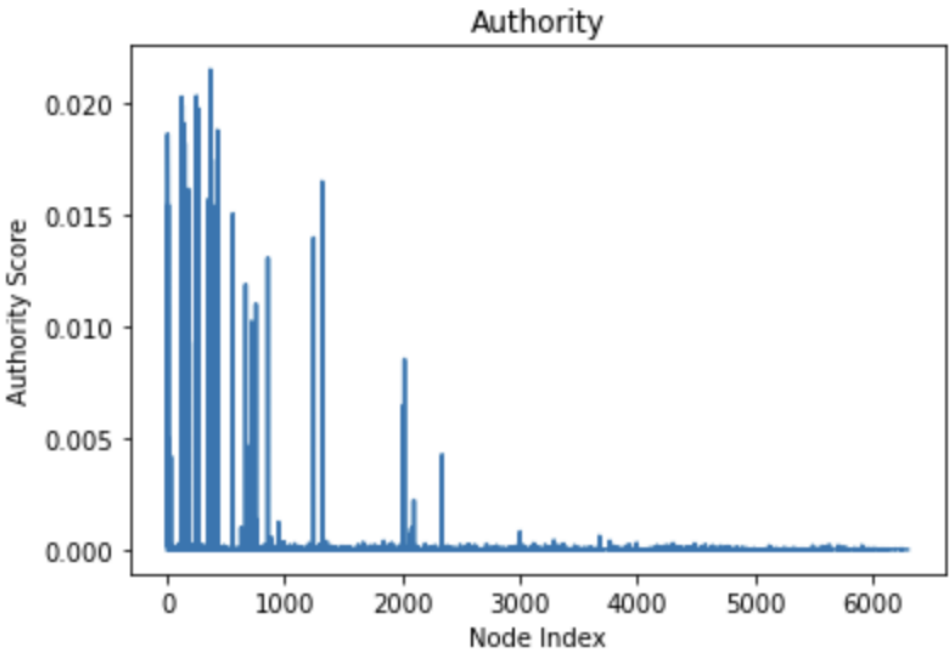
A node with high hub score shows that there are many nodes referred (high number of outgoing links) by the current node with high authority score thereby making it a perfect hub to find good authoritative web pages.

For example, Node 3 has high pagerank score which means its well connected and have many nodes directing towards it increasing its visibility. Also, these nodes have very high hub score which boosts the authority score of the Node 3 as well. Node 3 has 0 hub score which shows that either it has no outgoing links or its outgoing links have non-existent authority score.

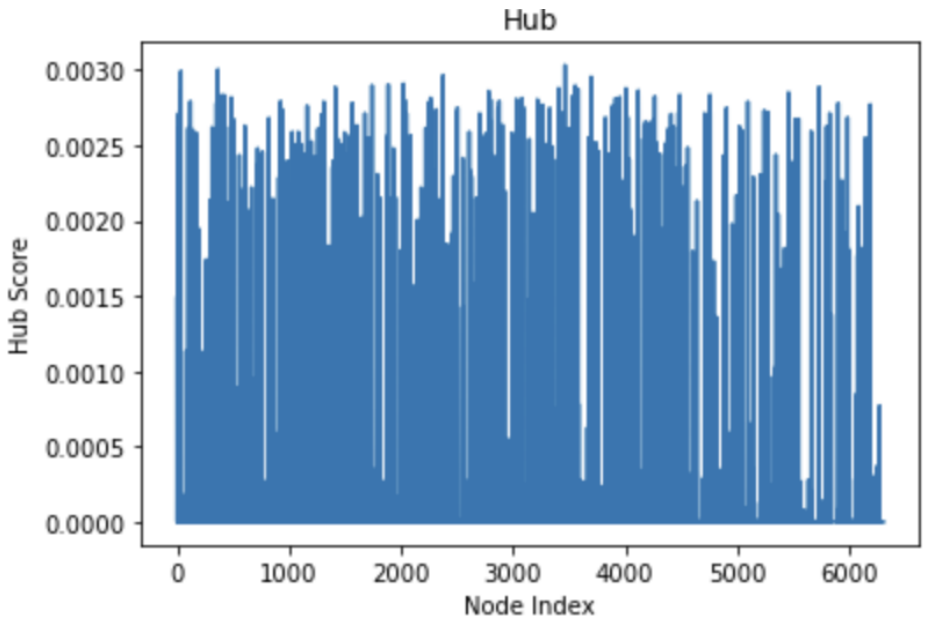
Plot of PageRank score for each node:



Plot of Authority score for each node:



Plot of Hub score for each node:


Hub
Hub Score
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
1000
2000
3000
4000
5000
6000
Node Index