IR ASSIGNMENT 3
GROUP 112

Question 1:

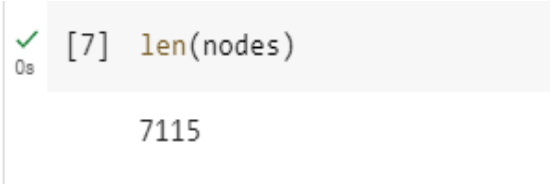
The wiki-vote dataset is a directed social network graph that represents Wikipedia users and their voting behavior The nodes in the graph represent Wikipedia users, while the edges represent votes made by one user for another. Wikipedia voting on promotion to administratorship (till January 2008). Directed edge A->B means user A voted on B becoming Wikipedia administrator. The dataset contains 7,115 nodes (Wikipedia users) and 103,689 edges (votes). It is publicly available on the Stanford Network Analysis Project (SNAP) website.

In the code part, the text file Wiki-Vote.txt is opened in read mode and each file is read line by line. Each row is a pair of two integers. The pair is stored as a tuple into a list called edges.

Adjacency Matrix :
Initially a matrix (size of no of nodes x no of nodes) is initialized with 0 entries. Then a mapping for converting the userIds to a nodeId is generated, along with a reverse map which will reverse map the index number of the node back to the user ids. Then we iterate through the edge list and store entries in cells adj_matrix[map[n1][map[n2]]=1

Edge List:
The adjacency list each list in the edge list represents the neighbors list of the nodes represented by the index of the list appending the second vertex to the list corresponding to the edge.

1. Number of Nodes- 7115
   Each node that is either casting a vote or being casted on in the election is stored in a set called ‘nodes’. The number of nodes is calculated by finding the size of the set nodes.



2. Number of Edges- 103689
   The number of edges is the number of tuples stored in edges.

```
[6]  len(edges)

     103689
```

3. Avg In-degree-
To determine the in-degree of a vertex, for this, we have to count the number of edges that end at the vertex, this is done for all the vertices, and in degrees are summed then divided by the total number of nodes.

4. Avg. Out-Degree-
The out-degree of a vertex can be calculated by finding the number of edges coming out from the node.Hence the length of the neighbor list corresponding to the node is stored in the out-deg array. Finally summation is done and divided by total nodes.

```
print("Average In Degree: ", avg_in)
print("Average Out Degree: ", avg_out)
```

```
Average In Degree:  14.573295853829936
Average Out Degree:  14.573295853829936
```

5. Node with Max In-degree
This can be interpreted as the node who is being pointed at the most number of times or with max number of votes.
For this, the in_deg array is traversed and we find the index with the maximum value of in degree. This is then reverse mapped to find the user id with the maximum number of votes.

6. Node with Max out-degree
This is the node who has cast the most number of votes or has the maximum value in out_deg array. The index is then reverse mapped to find the user id who voted the maximum number of times.

```
] print("The userID with max in-deg: ", reverse_map[node_in])
  print("The userID with max out-deg: ", reverse_map[node_out])
```

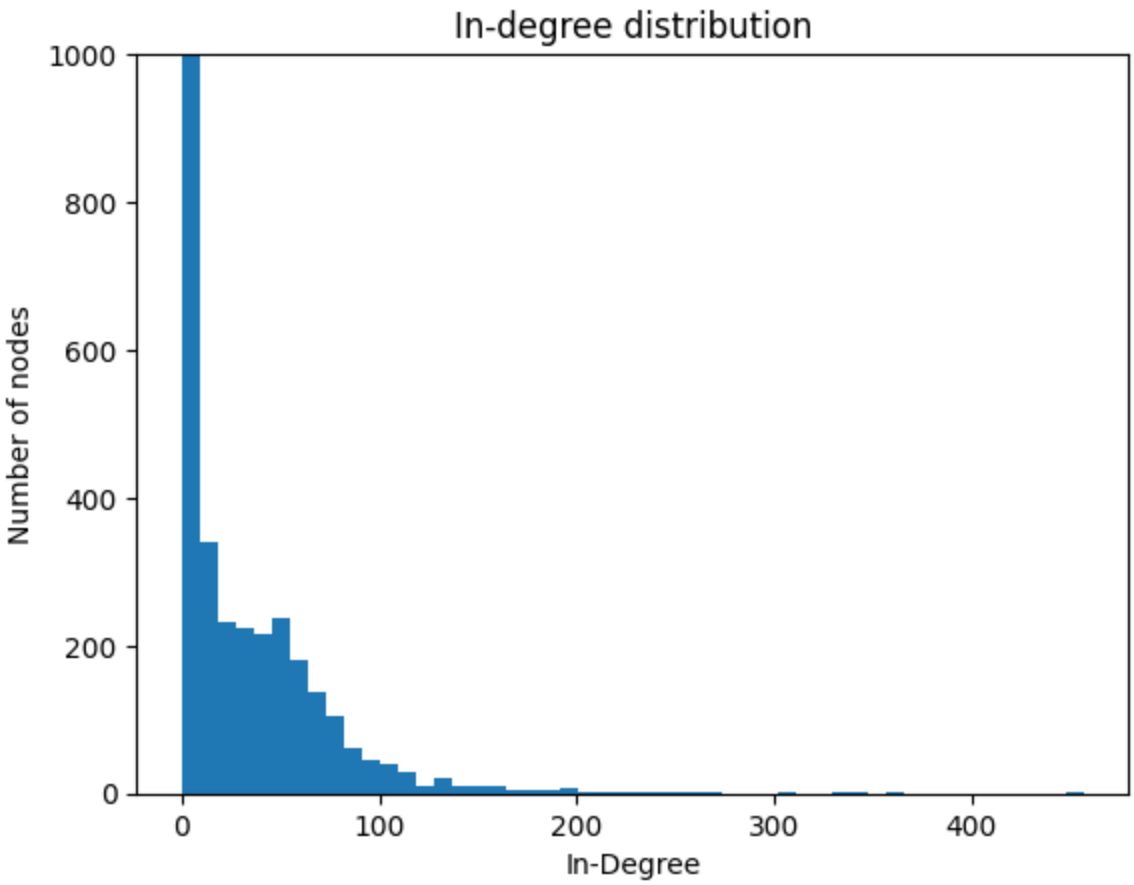
```
The userID with max in-deg:  4037
The userID with max out-deg:  2565
```

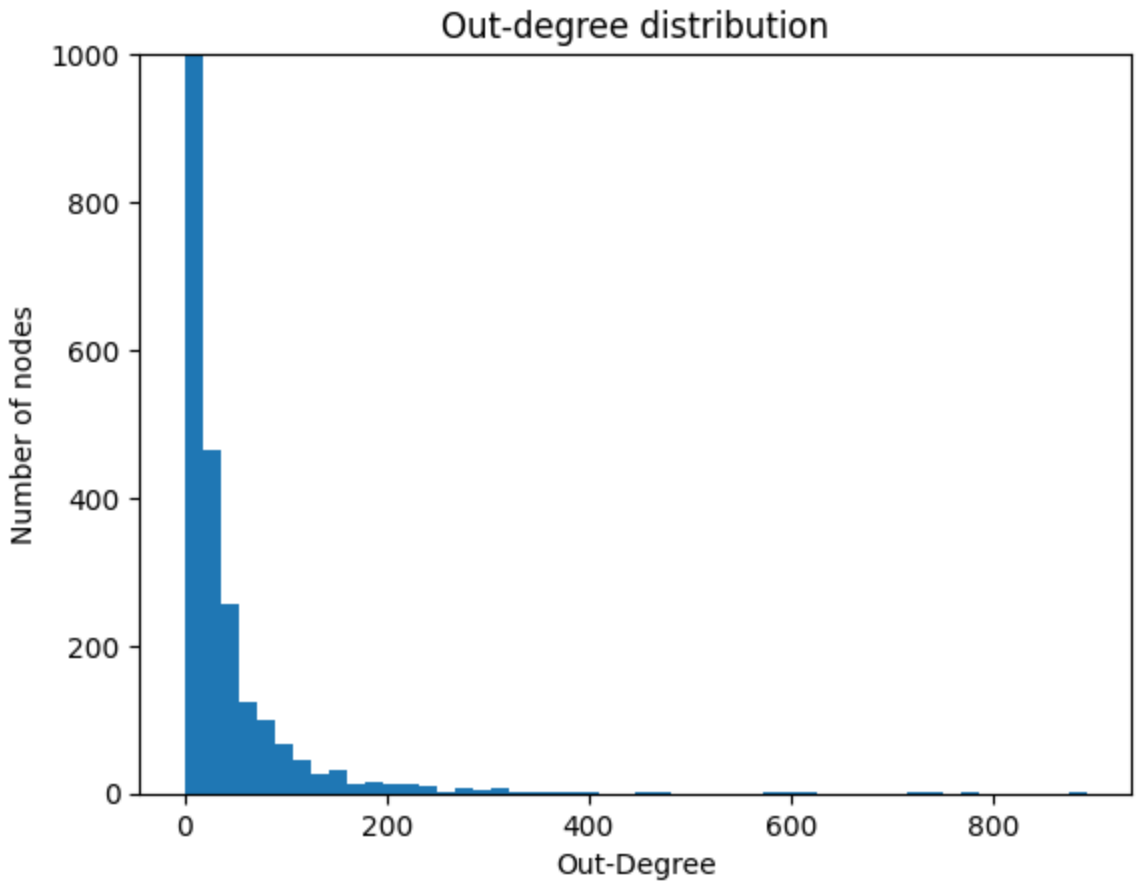
7. The density of the network
density = number of edges / maximum possible number of edges
max_edges=n*(n-1)

```
print("The density of the network is ", density)
```

```
The density of the network is  0.0020485375110809584
```

Degree Distribution of In Degree and Out-Degree
A histogram used to visualize the degree distribution of a network because it allows us to easily see the distribution of node degrees in the network. In a histogram, the x-axis represents the range of degrees, and the y-axis represents the number of nodes with that degree


In-degree distribution
Number of nodes
In-Degree
0
200
400
600
800
1000
0
100
200
300
400

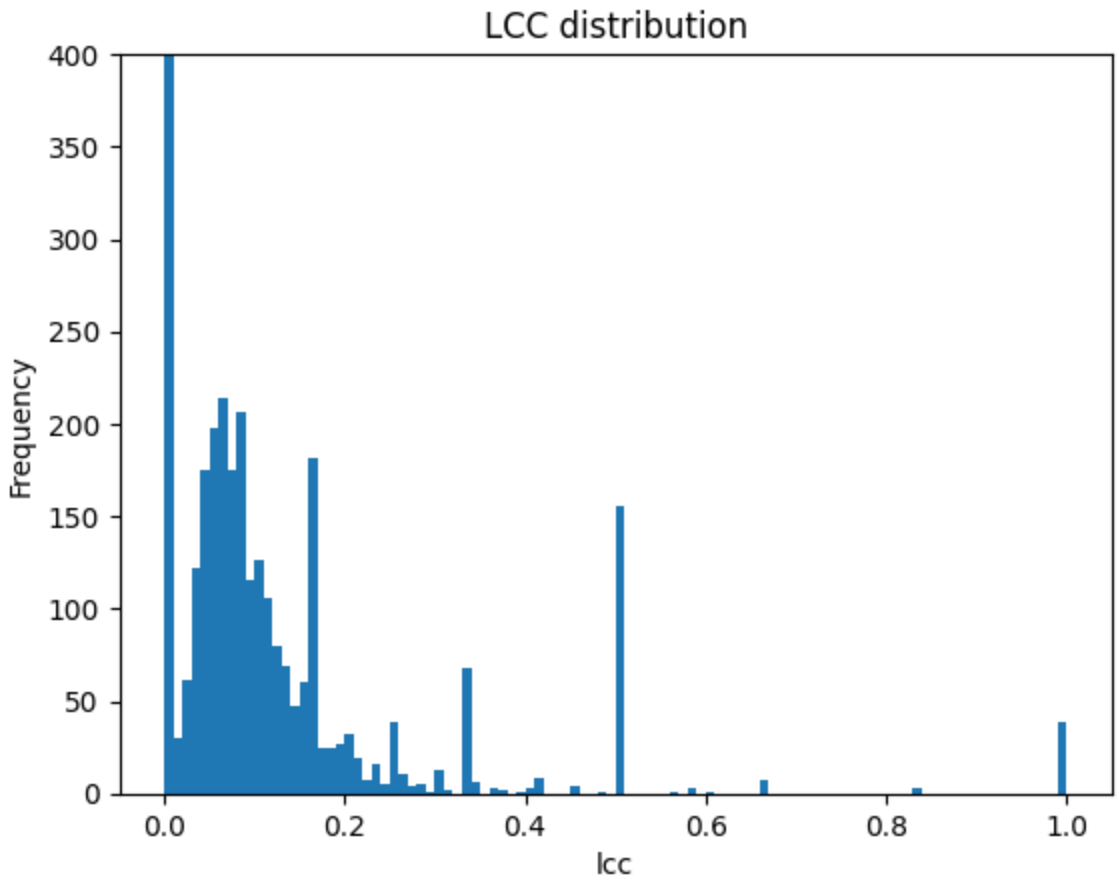


local clustering coefficient:
Lcc for a node can be calculated by dividing the number of edges between node's neighbors divided by the max number of possible edges that could have been between each of the neighbors, given by n*n-1 where n is the number of neighbors corresponding to the node.

$$LCCi = \frac{|(Ejk,\ j \text{ and } k \text{ are the neighbors of } i)|}{\text{Max edges bw the neighbors of } i}$$

First we extract the neighbors from the edge list of a node, then calc the max no of edges for the denominator. Then for numerator, for each neighbor in the edge list, the number rof edges it shares with other neighbors is calculated.

This ratio is found for each node. It is 0 if a node has no neighbor at all.


LCC distribution
Frequency
lcc
400
350
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0

.

Question 2 - [35 points] PageRank, Hubs and Authority

The PageRank, Hubs and Authority algorithms calculates a node's importance in a network based on the amount and quality of links pointing to it. It considers a node to be important if it is related to other important nodes. By recursively propagating the importance of nodes that point to it, the approach computes a score for each node. This makes PageRank appropriate for ranking websites in a hyperlink network, where pages connected to a large number of other key pages are deemed more important.

The program calculates the PageRank score of each node in a directed network using the `networkx` library in Python.

1. `import networkx as nx`: This imports the `networkx` library

2. `G = nx.DiGraph()`: This creates an instance of a directed graph (DiGraph) using the `networkx` library, which will be used to represent the directed network.

3. `list_edges = []`: This initializes an empty list to store the edges (links) between nodes (webpages) in the network.

**4. `for key, value in edge_list.items():** list_edges.append((key,value))`: This iterates through the `edge_list` dictionary, which presumably contains the edges (links) between nodes (webpages) in the directed network, and appends each edge as a tuple of `(key, value)` to the `list_edges` list.

5. `for node, neighbors in list_edges: G.add_node(node) ...`: This iterates through the `list_edges` list, and for each edge, adds the source node (`node`) and the target node (`neighbor`) to the directed graph `G` using the `add_node` and `add_edge` methods of `networkx`, respectively. This creates the directed network with nodes and edges.

6. `pagerank_scores = nx.pagerank(G, alpha=0.85)`: This calculates the PageRank scores of each node in the directed network `G` using the `pagerank` function of `networkx`, with a damping factor (alpha) of 0.85. The PageRank algorithm is a widely used algorithm that assigns a score to each node in a network based on the structure of the network, with higher scores indicating more important nodes.

7. Pagerank_scores are printed. For example:

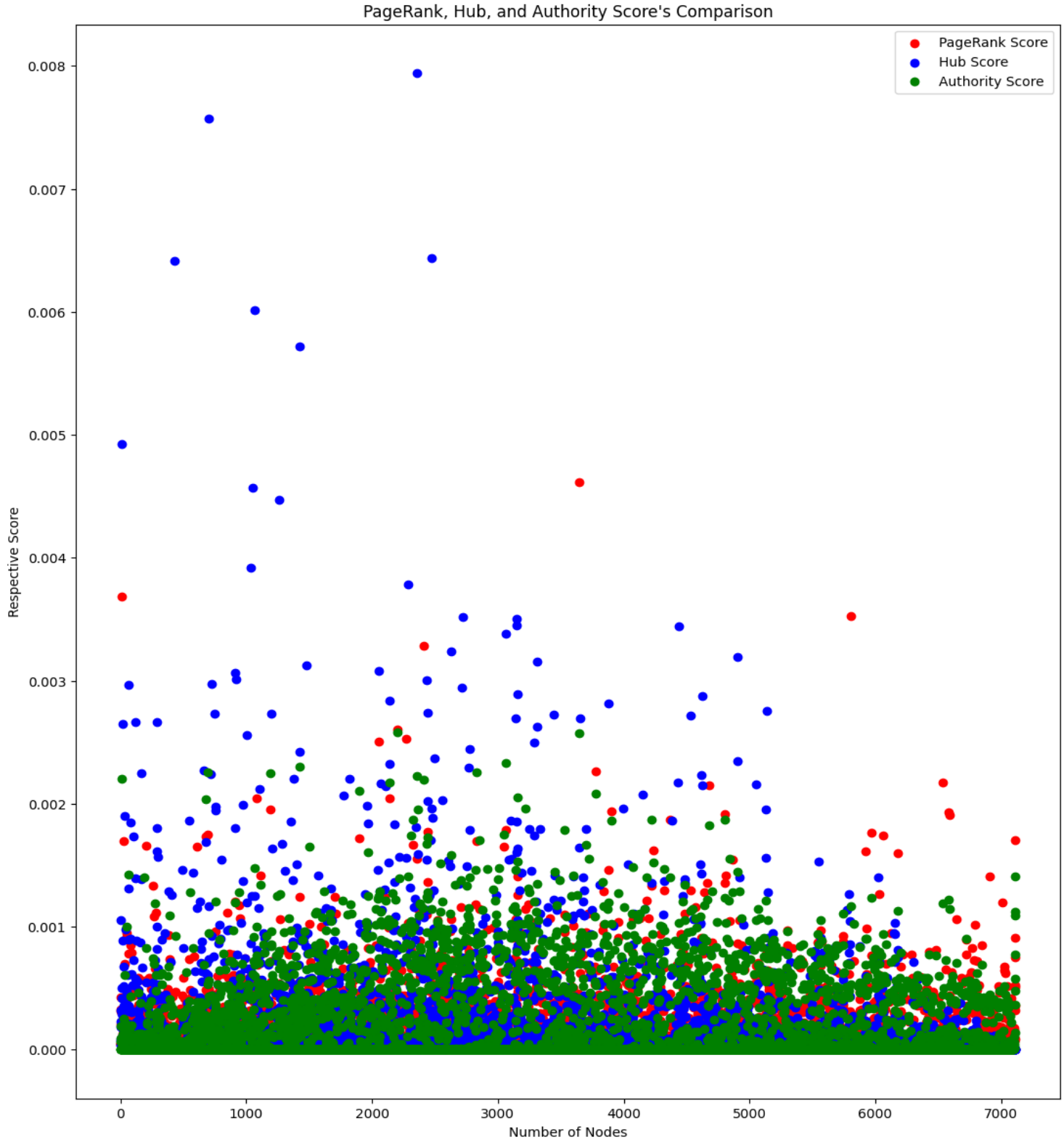
```
   435 :  5.0487823458630175e-05
   436 :  5.0487823458630175e-05
   3569 :  0.00010242162029977101
   3571 :  0.00016624346374375038
   438 :  5.0487823458630175e-05
   439 :  5.0487823458630175e-05
   156 :  5.0487823458630175e-05
   441 :  5.0487823458630175e-05
   443 :  5.0487823458630175e-05
   445 :  5.0487823458630175e-05
   446 :  5.0487823458630175e-05
   447 :  5.0487823458630175e-05
   448 :  5.0487823458630175e-05
   442 :  5.0487823458630175e-05
   450 :  5.0487823458630175e-05
   451 :  5.0487823458630175e-05
   452 :  5.0487823458630175e-05
   453 :  5.0487823458630175e-05
   454 :  5.0487823458630175e-05
   456 :  5.0487823458630175e-05
   457 :  5.0487823458630175e-05
   458 :  5.0487823458630175e-05
   459 :  5.0487823458630175e-05
   460 :  5.0487823458630175e-05
   461 :  5.0487823458630175e-05
   463 :  5.0487823458630175e-05
   466 :  5.0487823458630175e-05
   465 :  5.0487823458630175e-05
```

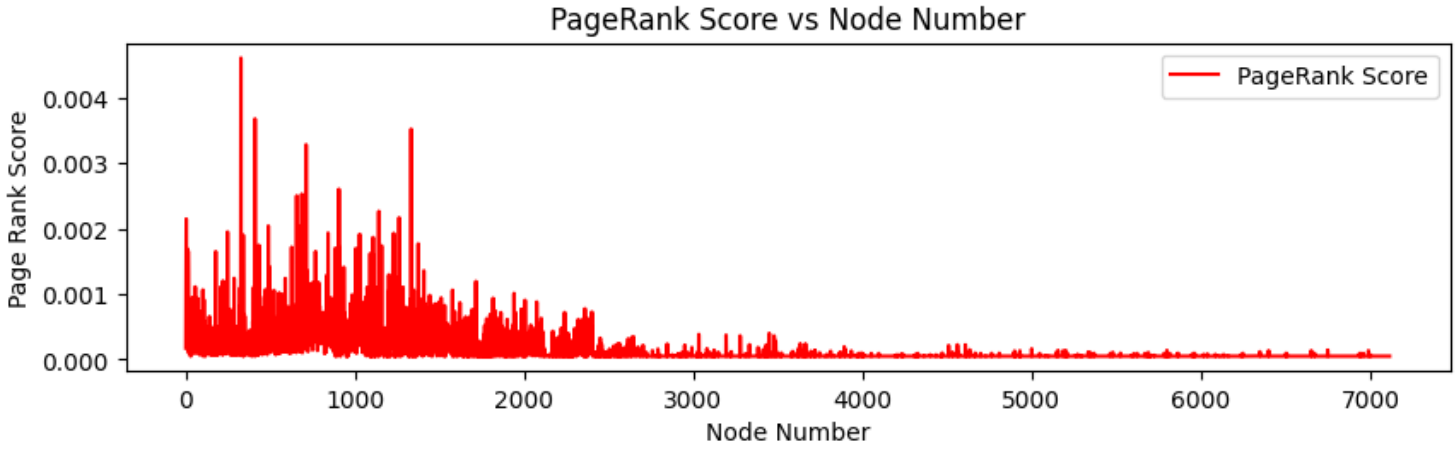
8. h, a = nx.hits(G): Here we used the hits algorithm function, which in return gives dictionaries, h and a, storing the key values of hub score and authority score of each node, respectively. This HITS algorithm basically selects the nodes that are good hubs or authority based on their relationships with other nodes in the network. HITS assigns two ratings to each node, one for its Authority value and one for its hub value. A good authority is usually the one which is pointed to by many other good hubs, while, on the other hand, a good hub is one that points to many other good authorities. HITS may be used to rank websites that provide good answers to inquiries, where a query node can be linked to numerous web pages.

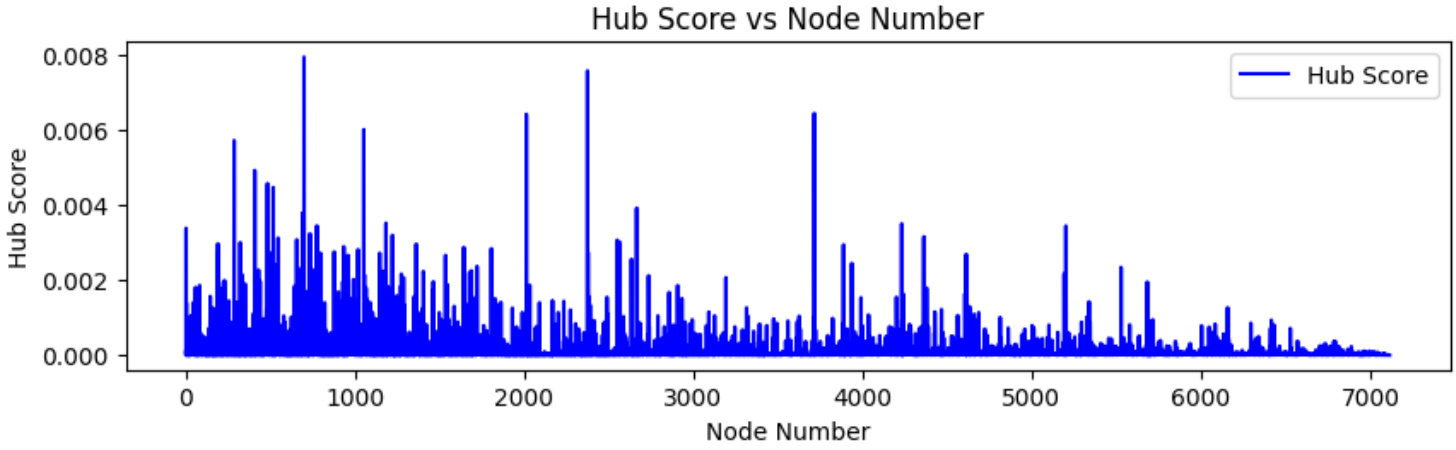
```
Node Number  6951 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  6.338684346439843e-05  &  Authority=  -3.2297195147388e-20
Node Number  6952 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  1.684394319872553e-05  &  Authority=  3.8513286155897524e-22
Node Number  6953 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  1.879810365720256e-05  &  Authority=  -2.955189423361182e-20
Node Number  6954 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  1.684394319872553e-05  &  Authority=  2.024489664765283e-21
Node Number  6955 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  1.684394319872553e-05  &  Authority=  1.2336285120982072e-22
Node Number  6957 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  -1.5684669985573674e-20
Node Number  6958 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  -3.3774583413383024e-21
Node Number  6959 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  -2.2821310970519268e-20
Node Number  6960 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  8.086804814708911e-20
Node Number  6961 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  2.22900855596248e-20
Node Number  6967 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  -3.1509975679827086e-22
Node Number  6962 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  1.811645794127426e-20
Node Number  6963 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  5.806883972568278e-22
Node Number  6964 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  -1.9493173809909983e-20
Node Number  6970 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  4.75113018249495e-05  &  Authority=  -3.517866642878105e-20
Node Number  6971 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.69342233704893e-05  &  Authority=  -9.598183291625563e-21
Node Number  6965 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  -5.556435406325176e-22
Node Number  6966 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  0.00010564384615798048  &  Authority=  -3.881501998309837e-21
Node Number  6972 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  3.1711341793563665e-21
Node Number  6973 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  1.1946075506710045e-20
Node Number  6968 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  -1.2995400039293988e-20
Node Number  6969 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.225503762599401e-05  &  Authority=  1.6581624355604764e-20
Node Number  6974 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  4.826355771028636e-06  &  Authority=  2.409606270552714e-21
Node Number  6975 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  4.826355771028636e-06  &  Authority=  1.8666113258201457e-20
Node Number  6976 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  4.826355771028636e-06  &  Authority=  3.3320332099181e-20
Node Number  6980 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  4.291711968224537e-06  &  Authority=  -1.6957779333063125e-21
Node Number  6983 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  2.7431834637304224e-06  &  Authority=  5.597059971041243e-21
Node Number  6985 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  1.96817235665659e-06  &  Authority=  -1.5714597330492528e-20
Node Number  6987 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  5.956817239328042e-06  &  Authority=  3.084382838114856e-22
Node Number  6990 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  5.956817239328042e-06  &  Authority=  1.0250012302583836e-21
Node Number  6988 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  5.956817239328042e-06  &  Authority=  4.7479351310325696e-20
Node Number  6956 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  2.1213081438283538e-05  &  Authority=  -7.39840776156782e-20
Node Number  6989 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  5.956817239328042e-06  &  Authority=  2.5460894452498178e-20
Node Number  6992 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  4.679185744495289e-06  &  Authority=  5.756444488344545e-20
Node Number  6994 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  4.679185744495289e-06  &  Authority=  1.2963888584231516e-20
Node Number  6995 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  4.679185744495289e-06  &  Authority=  3.4498202416604996e-20
Node Number  6996 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  6.06278328657751e-06  &  Authority=  -3.898439702604182e-21
Node Number  6997 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  5.58435905459748e-06  &  Authority=  -2.2462156773234813e-20
Node Number  6998 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  1.1273854736576395e-06  &  Authority=  -2.981550199136833e-21
Node Number  7002 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  8.493979064031558e-05  &  Authority=  1.969867123766847e-20
Node Number  7003 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  8.702095236037742e-06  &  Authority=  1.175099418043686e-20
Node Number  7004 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  1.5256264198955495e-05  &  Authority=  -1.806939307058064e-20
Node Number  7005 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  3.7145853673082166e-05  &  Authority=  -1.823103148196676e-21
Node Number  7007 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  1.5256264198955495e-05  &  Authority=  -9.855735457043306e-21
Node Number  7008 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  1.5256264198955495e-05  &  Authority=  -2.557397165160834 3e-20
Node Number  7006 :  Page Rank Score=  5.0487823458630175e-05 ,  Hub=  2.99021154398897e-05  &  Authority=  -2.0668403848243767e-21
```

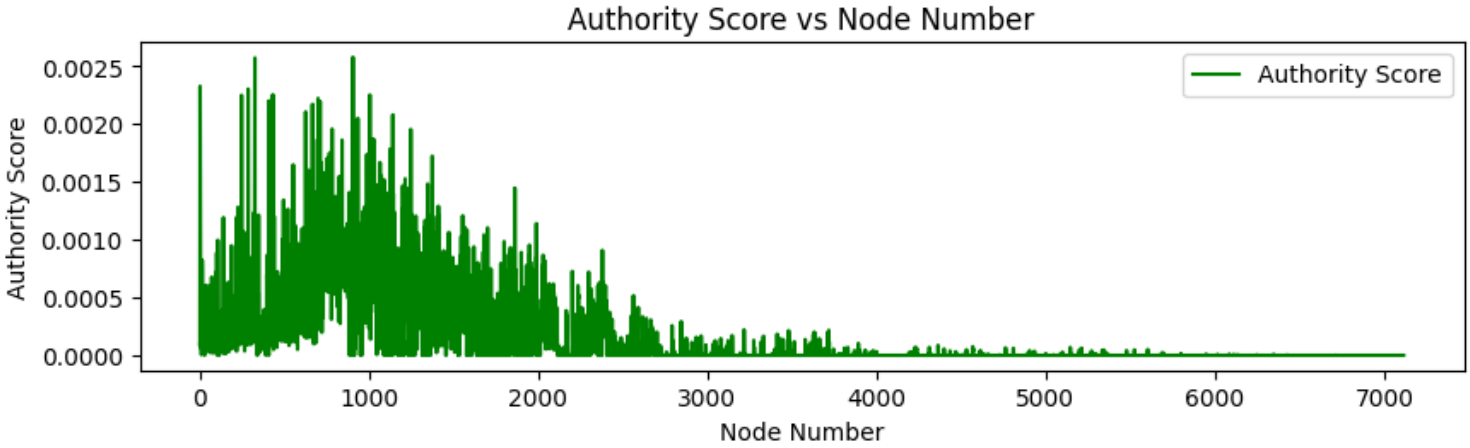
Comparison -

While the two mechanisms use different scoring techniques, they are related in that a node's PageRank score is influenced by its HITS authority score, and a node's HITS hub score is influenced by its PageRank score. In other words, a high HITS authority score will almost certainly be a high PageRank score, and a high PageRank score will almost certainly be a high HITS hub score. As a result, both approaches may be used in tandem to generate a more comprehensive score of network nodes.







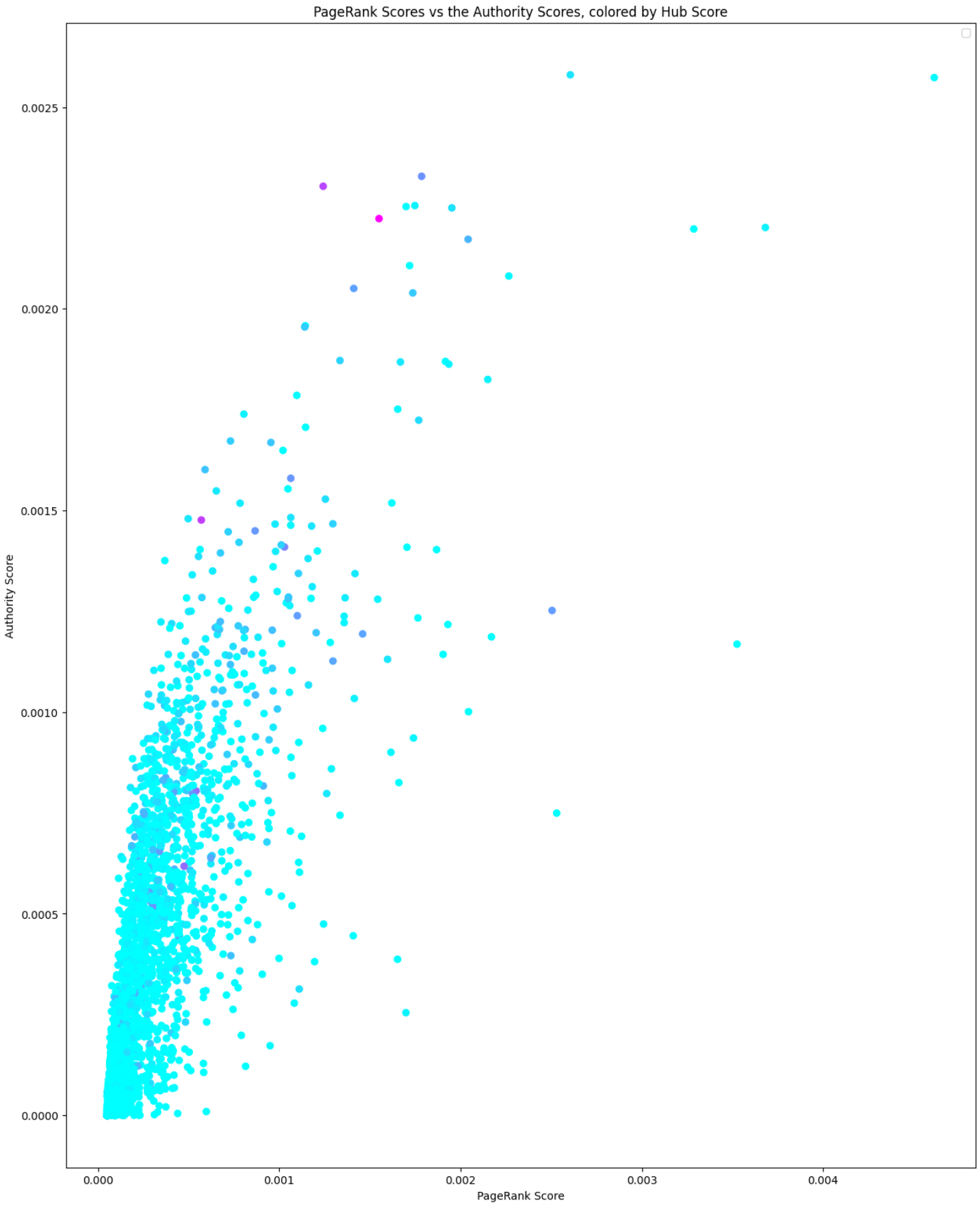


In the scatter plot below, we have the PageRank scores on the x-axis and the respective Authority scores on the y-axis, and eventually colored by Hub scores. If the **PageRank and Authority scores have a high positive correlation** (in this case here, refer to the figure below), the dots in the plot will tend to form a diagonal line from the bottom left corner to the upper right corner. This indicates that nodes with high PageRank tend to have a relatively high Authority, and this goes the back way as well, that is, the nodes with high authority scores and a high page rank score as well!

Similarly, if the Hub and Authority scores have a significant positive correlation, the dots in the plot will tend to form a diagonal line but now from the upper left corner to the lower right corner in this case here. As a result, nodes with high Hub scores also have high Authority scores, and this goes the back way as well, that is, the nodes with high authority scores and a high hub rank score as well.

If there is no dependency/ correlation between the three scores here, the plot's points will be dispersed on a random basis across the complete graph/ plot. Thus, implying that the three scores of the nodes (Page rank, Authority, and Hub scores) are not strongly related to one another.

By looking at the scatter plot and the patterns formed by the points are shown below:


PageRank Scores vs the Authority Scores, colored by Hub Score
Authority Score
PageRank Score
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.000
0.001
0.002
0.003
0.004

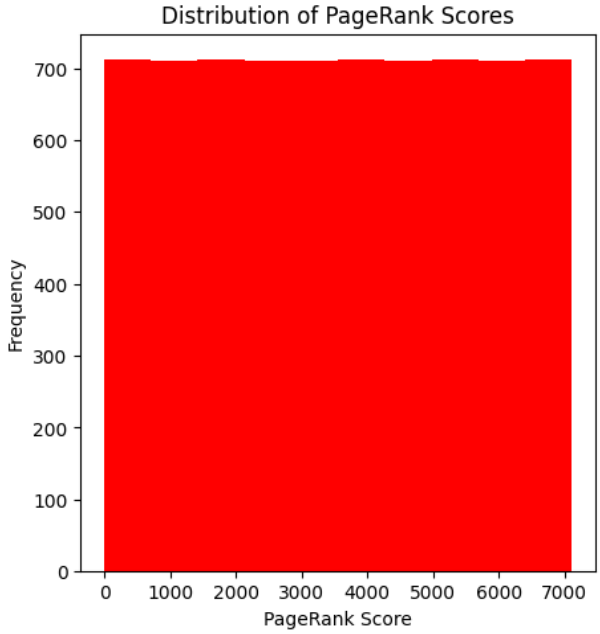
Distribution of different scores across different frequencies:

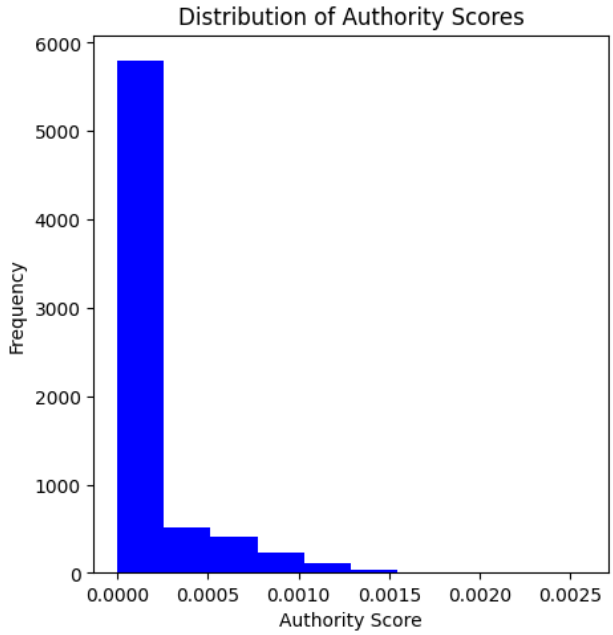
The three histograms show the distribution of the PageRank, Authority, and Hub scores across all nodes in the network. The x-axis of each histogram represents the different scores (Pagerank, Hub, Authority), and the y-axis represents the frequency of nodes with that respective score.

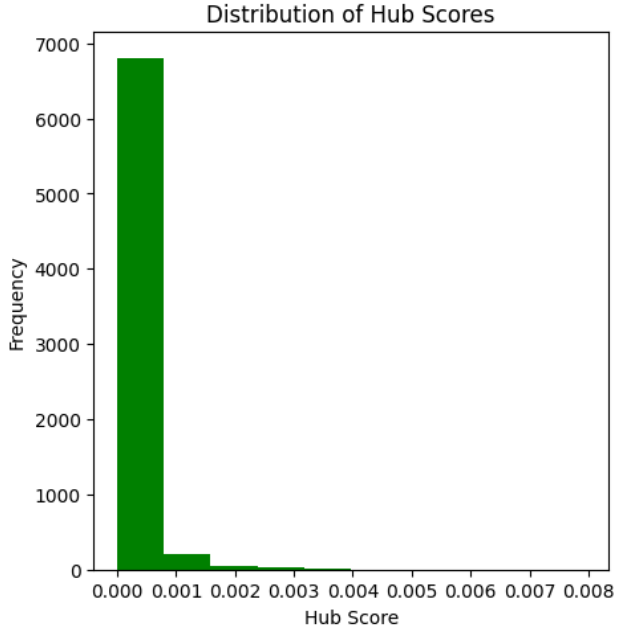
The shape of each histogram shows how the scores are distributed across the nodes in the network, and important analysis can be derived from the same. Let's say that if a histogram is relatively flat shaped, that is: more evenly distributed **(here: Pagerank score's histogram)**, this means that the Authority scores are more evenly distributed across the nodes in the network that we have. Similarly, if a histogram is somewhat skewed to the left, like with a long tail on the left-hand side **(here: Authority and Hub score's histogram),** this means that there are a few nodes with very high Authority and Hub's scores, respectively, while the majority of nodes have relatively low scores which is quite evident from the graph and explains its skewed look. Here the Authority score histogram is more evenly distributed as compared to the Hub score one since the majority of Hub score frequencies are concentrated on the lower Hub score section, making it more skewed.

The histograms basically offer a better understanding of the network's features. It gives the insights to the nodes with the highest influences relatively, nodes which are closely connected and which are far apart, and maybe also the nodes that serve as intermediaries or links and basically the bridges between the various sections of the network that we have. This also gives a good idea of the broad distribution of the scores across different frequncies.

Please find the graph below:


Distribution of PageRank Scores
Frequency
PageRank Score
0
100
200
300
400
500
600
700
0
1000
2000
3000
4000
5000
6000
7000


Distribution of Authority Scores
Frequency
Authority Score
0
1000
2000
3000
4000
5000
6000
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025


Distribution of Hub Scores
Frequency
Hub Score
0
1000
2000
3000
4000
5000
6000
7000
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008