Report

Question 1

We were asked to pick a real-world network dataset with nodes > 100, so for this purpose, we have chosen wiki-vote to be our dataset. The network has information about Wikipedia voting (related to the promotion of users to adminship) and the data is collected from the period of the start of Wikipedia to January 2008. The nodes in the network represent the Wikipedia users, and a directed edge in the graph from i th node to j th node represents that user i voted for user j.

Methodology:

Firstly, it was observed that the nodeIDs were not continuous in a particular range of integers thus we assigned a unique node number to each node and maintained a map from nodeID to nodeNumber and vice versa too. The nodeNumber is an integer in the range [0, N] where N = Number of nodes in the graph.
Following this, we form an ‘adjacency matrix’ and ‘edge list’ representing the network. The ‘adjacency matrix’ A is an NxN matrix where Aij=1 if there is an edge from the node with nodeNumber ‘i’ to the node with nodeNumber ‘j’ in the directed network otherwise, Aij = 0. Edge list is a list of tuples where each tuple (i,j) represents that there is an edge from the node with nodeNumber ‘i’ to the node with nodeNumber ‘j’ in the directed network.

Following is the information about the dataset:
1. Number of Nodes: 7115
2. Number of Edges: 103689
3. Avg In-degree: 14.573295853829936
4. Avg. Out-Degree: 14.573295853829936
5. Node with Max In-degree : Node ID - 4037 | In-Degree = 457.0
6. Node with Max out-degree : Node ID - 2565 | Out-Degree = 893.0
7. The density of the network: 0.0020485375110809584

Metrics calculation:

- The number of nodes and number of edges were directly present in the dataset file.

- Avg In-degree: Firstly indegree of all nodes is calculated and stored in a form of a dictionary with the key representing the node number and value representing the in-degree of that node. For calculating the in-degree of a node with node number ‘i’, we take the sum of elements of i th column because that would have value 1 whenever there is an edge incoming to node ‘i’ from any other node. After getting the in-degree of all the nodes, we simply take the average of the in-degree values over all nodes.

Average in-degree = $\frac{Sum\ of\ in-degrees\ of\ all\ nodes}{N}$

- Avg out-degree: Firstly the out-degree of all nodes is calculated and stored in a form of a dictionary with the key representing the node number and the value representing the out-degree of that node. For calculating the out-degree of a node with node number ‘i’, we take the sum of elements of i th row because that would have value 1 whenever there is an edge outgoing from node ‘i’ to any other node. After getting the out-degree of all the nodes, we simply take the average of the out-degree values over all nodes.

Average out-degree = $\frac{Sum\ of\ out-degrees\ of\ all\ nodes}{N}$

- We calculated the in-degrees of all the nodes previously. So we select the node with the maximum value of indegree from the dictionary of indegree we have maintained.

- We calculated the out-degrees of all the nodes previously. So we select the node with the maximum value of out-degree from the dictionary of out-degree we have maintained.

- For network density we have used the formula :

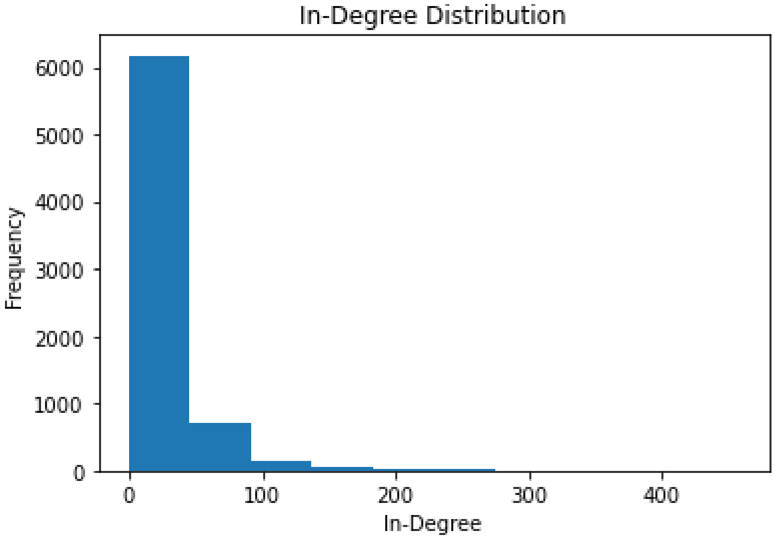
Network Density = $\frac{No.\ of\ edges\ present\ in\ the\ network}{Total\ possible\ edges\ in\ the\ network}$

In a directed graph network,

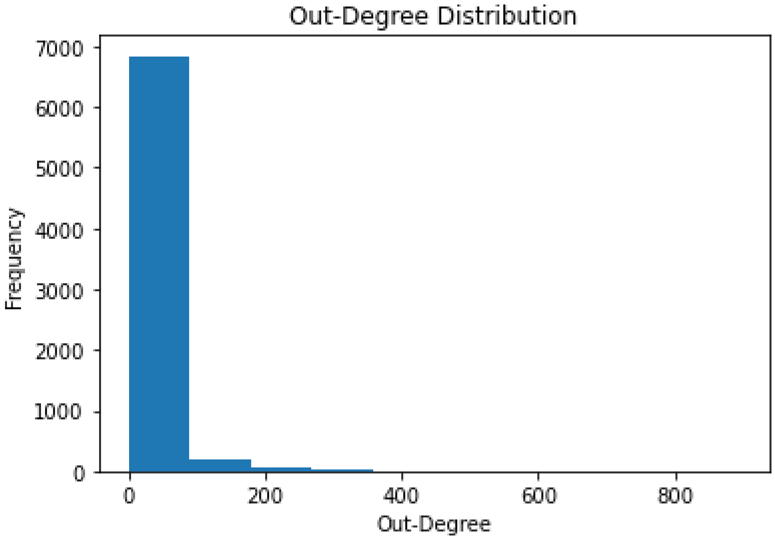
Total possible edges in network = (N)*(N - 1)
where, N = Number of nodes in the graph

Following are the degree distribution plots for the network:

- In-Degree Distribution:



- Out-Degree Distribution:



Local Clustering Coefficient (LCC):

Note: The graph is converted to an undirected version as instructed for this part.

The approach followed for calculating the LCC for each node is as follows:

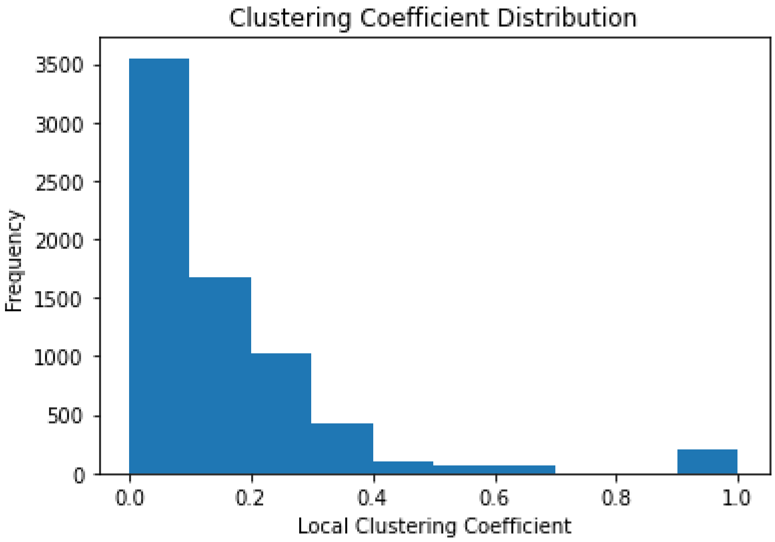
For a node, we first find the neighbourhood i.e the set of all the neighbour nodes that the given node is directly connected to. Let the number of neighbouring nodes for node ‘i’ be NV_i. We then find the number of edges(links) existing between the neighbours of ‘i’, let it be NE_i. Then the value of LCC for node ‘i’ is :

Total possible links between neighbours = NT_i = (NV_i)*(NV_i - 1) / 2
[because we have considered undirected version here]

LCCi = NE_i / NT_i

This way, the local clustering coefficient of each node is calculated.

The plot for clustering-coefficient distribution is :



Question 2

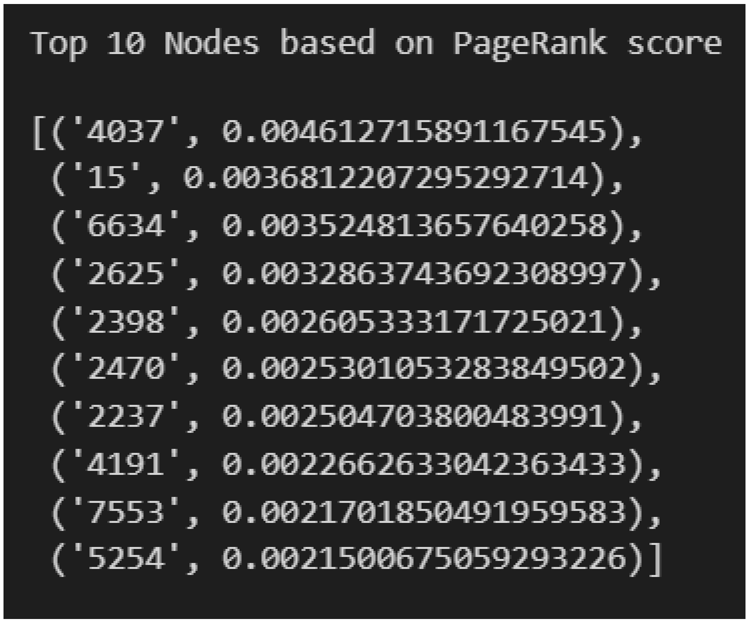
Calculating the page rank score for each node.

We first read the contents of the file named "Wiki-Vote.txt" using the 'with' statement and storing it in the 'text' variable. We then split the contents of the file based on the '#' symbol and store the resulting list of strings in the 'lst' variable.

Next, we extract the edges of the graph from the third element of the 'lst' list and split it into a list of edges using the '\n' and '\t' characters as delimiters. We store this list of edges in the 'edges' variable.

We then create a directed graph object 'G' using the edges extracted from the file using the NetworkX library. We calculate the PageRank score for each node in the graph using the 'pagerank' function provided by NetworkX and store the results in the 'pr' variable.

For ease, we have only displayed the Top 10 Nodes based on the PageRank score.



```
Top 10 Nodes based on PageRank score

[('4037', 0.004612715891167545),
 ('15', 0.0036812207295292714),
 ('6634', 0.003524813657640258),
 ('2625', 0.0032863743692308997),
 ('2398', 0.002605333171725021),
 ('2470', 0.0025301053283849502),
 ('2237', 0.002504703800483991),
 ('4191', 0.0022662633042363433),
 ('7553', 0.0021701850491959583),
 ('5254', 0.0021500675059293226)]
```

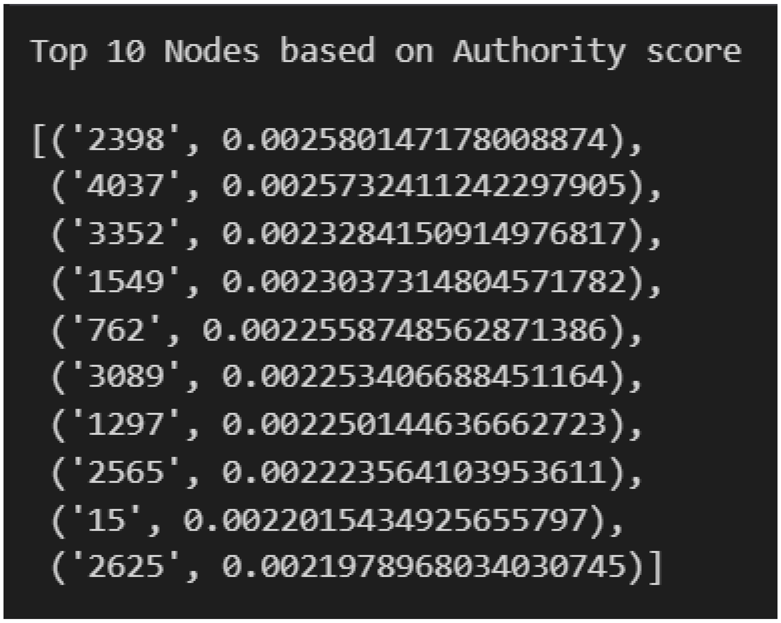
Calculating the authority and hub score for each node.

We used the HITS algorithm provided by the NetworkX library to calculate the hub and authority scores for each node in Graph 'G.' Authorities are those nodes that contain

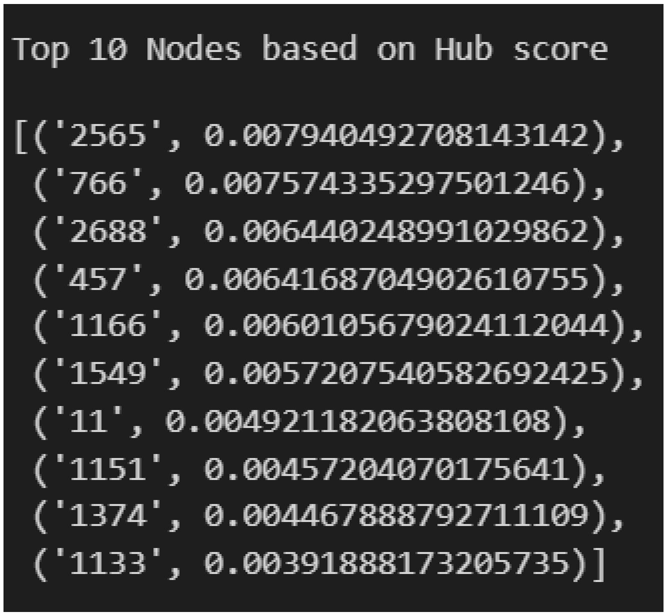
helpful information, and its importance is measured by incoming links, whereas hubs are the nodes that point toward authorities.
In mathematical terms, the authority score of a node X is the sum of hub scores of all the nodes that point to X. Hub score of a node X is the sum of authority scores of all the nodes that X points towards.

The 'max_iter' parameter specifies the maximum number of iterations used by the HITS algorithm while calculating the scores. The 'normalized' parameter specifies whether the scores should be normalized after each iteration.

For ease, we only display the top 10 Nodes having the highest score. To achieve this we sort the hub and authority scores for all nodes in descending order using the 'sorted' function and store the top 10 nodes with the highest hub and authority scores in the 'top_hubs' and 'top_auth' variables, respectively.

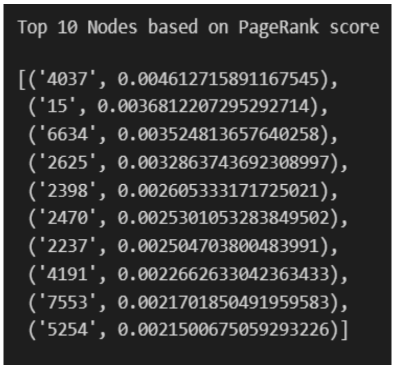
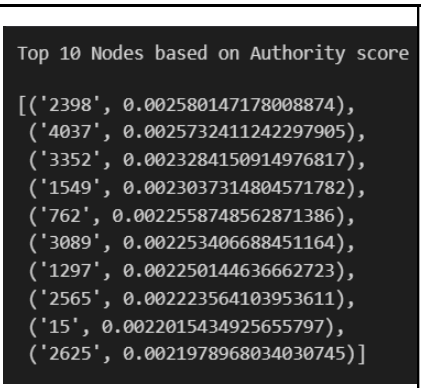
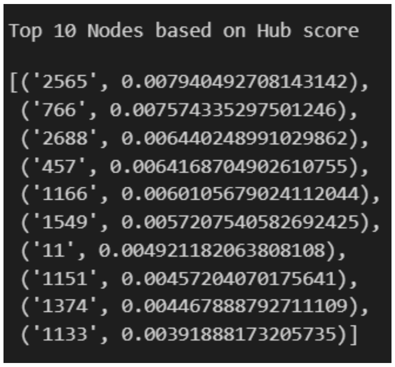


```
Top 10 Nodes based on Authority score

[('2398', 0.002580147178008874),
 ('4037', 0.0025732411242297905),
 ('3352', 0.0023284150914976817),
 ('1549', 0.0023037314804571782),
 ('762', 0.0022558748562871386),
 ('3089', 0.002253406688451164),
 ('1297', 0.002250144636662723),
 ('2565', 0.002223564103953611),
 ('15', 0.0022015434925655797),
 ('2625', 0.0021978968034030745)]
```

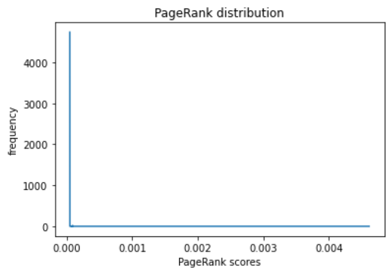


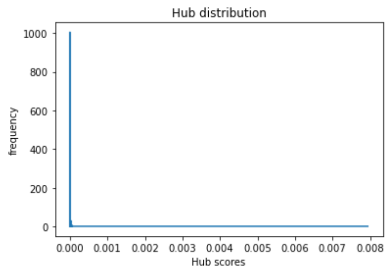
```
Top 10 Nodes based on Hub score

[('2565', 0.007940492708143142),
 ('766', 0.007574335297501246),
 ('2688', 0.006440248991029862),
 ('457', 0.0064168704902610755),
 ('1166', 0.0060105679024112044),
 ('1549', 0.0057207540582692425),
 ('11', 0.004921182063808108),
 ('1151', 0.00457204070175641),
 ('1374', 0.004467888792711109),
 ('1133', 0.00391888173205735)]
```

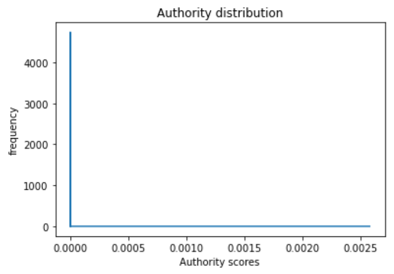
<u>Comparing the results obtained from both the algorithms in part 1 and part 2 based on the node scores.</u>

 Top 10 Nodes based on PageRank score [('4037', 0.004612715891167545), ('15', 0.0036812207295292714), ('6634', 0.003524813657640258), ('2625', 0.0032863743692308997), ('2398', 0.002605333171725021), ('2470', 0.0025301053283849502), ('2237', 0.002504703800483991), ('4191', 0.0022662633042363433), ('7553', 0.0021701850491959583), ('5254', 0.0021500675059293226)]	 Top 10 Nodes based on Authority score [('2398', 0.002580147178008874), ('4037', 0.0025732411242297905), ('3352', 0.0023284150914976817), ('1549', 0.0023037314804571782), ('762', 0.0022558748562871386), ('3089', 0.002253406688451164), ('1297', 0.002250144636662723), ('2565', 0.002223564103953611), ('15', 0.0022015434925655797), ('2625', 0.0021978968034030745)]	 Top 10 Nodes based on Hub score [('2565', 0.007940492708143142), ('766', 0.007574335297501246), ('2688', 0.006440248991029862), ('457', 0.0064168704902610755), ('1166', 0.0060105679024112044), ('1549', 0.0057207540582692425), ('11', 0.004921182063808108), ('1151', 0.00457204070175641), ('1374', 0.004467888792711109), ('1133', 0.00391888173205735)]

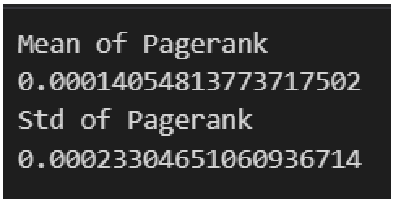
Since PageRank computes a ranking of nodes in the graph based on the structure of the incoming links and the same is the case while computing the authority score using the HITS algorithm, therefore PageRank and Authority scores can be similar. Therefore we can observe a lot of common nodeIDs in the top 10 based on PageRank and Authority score for eg - 4037, 15, 2625, 2398.
Also, we can see that Node with Max In-Degree: 4037 with In-Degree: 457 has a very high PageRank and Authority score.
Since the hub score is based on the outgoing links, therefore Node with Max Out-Degree: 2565 with Out-Degree: 893 has the highest Hub score.



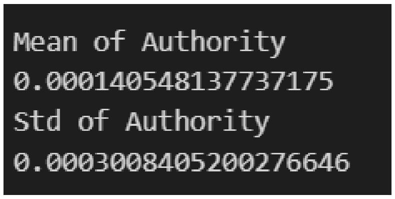




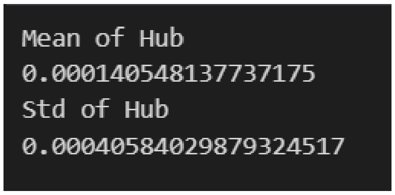
From the above three graphs, we can see that all three distributions have certain outlier nodes that have a significantly high score when compared to other nodes. The distribution is very dense around certain values which suggests that apart from the outlier nodes all other nodes tend to have a similar score for a particular distribution.



```
Mean of Pagerank
0.00014054813773717502
Std of Pagerank
0.00023304651060936714
```



```
Mean of Authority
0.000140548137737175
Std of Authority
0.0003008405200276646
```



```
Mean of Hub
0.000140548137737175
Std of Hub
0.00040584029879324517
```

We can see that the mean of all the scores is almost the same, but the standard deviation differs.