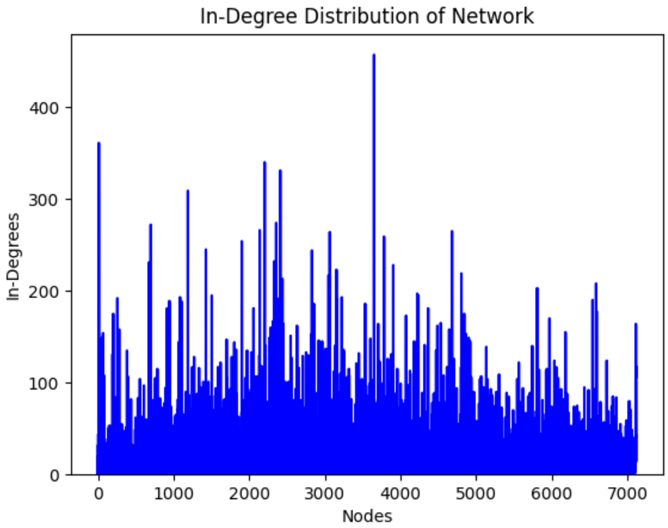
Information Retrieval
Assignment 3

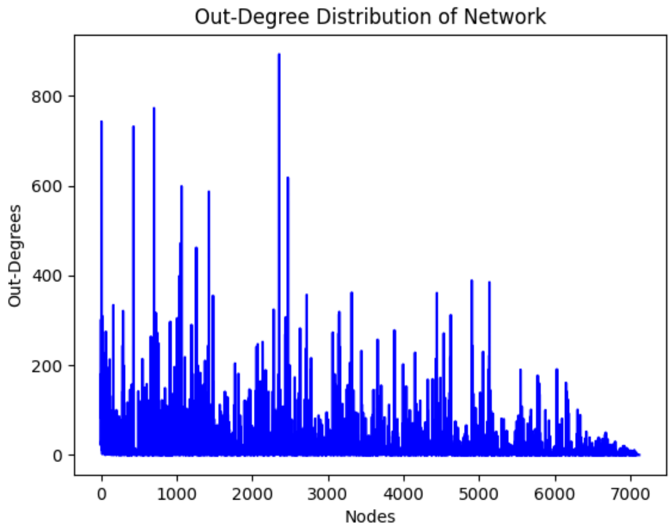
The dataset is on Wikipedia's voting on promotion to administratorship (till January 2008). It represents a directed graph where an edge A->B means user A voted on user B becoming a Wikipedia administrator.

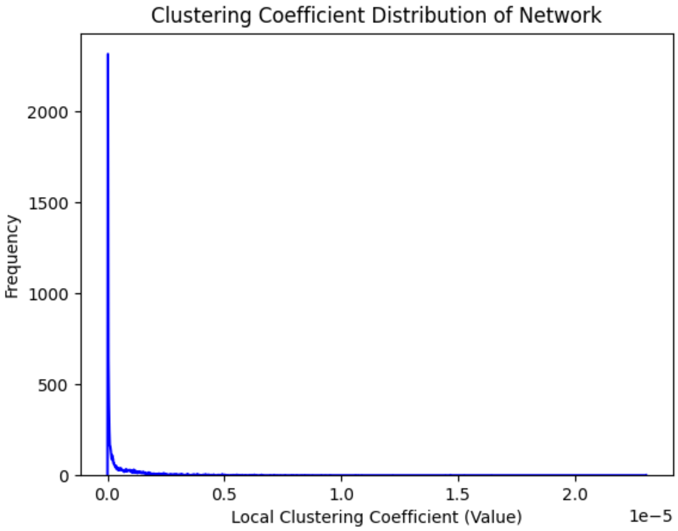
We first read the dataset, determine the number of unique nodes and assign each node number a unique index. We then create an adjacency list of the graph so obtained in order to perform analysis on it

Q1.

The following are the network details:
Nodes: 7115
Edges: 103689
Avg In-Degree: 14.573295853829936
Max In-Degree: 457
Max In-Degree Node: 3649 (Node Value - 4037)
Avg Out-Degree: 14.573295853829936
Max Out-Degree: 893
Max Out-Degree Node: 2356 (Node Value - 2565)
Density of the Network: 0.0020485375110809584





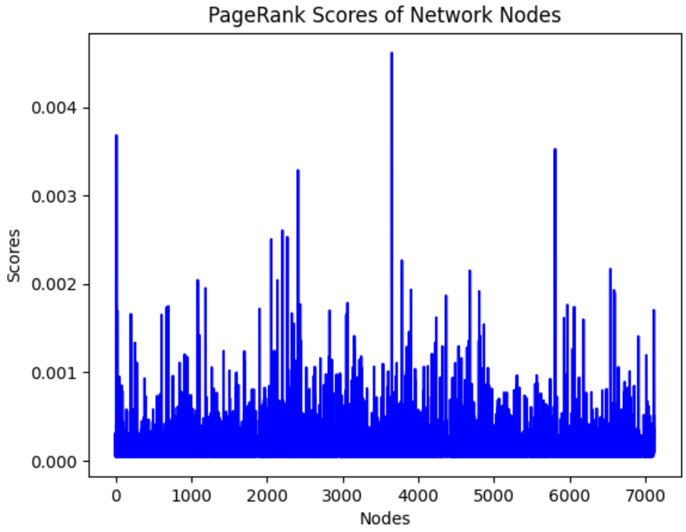


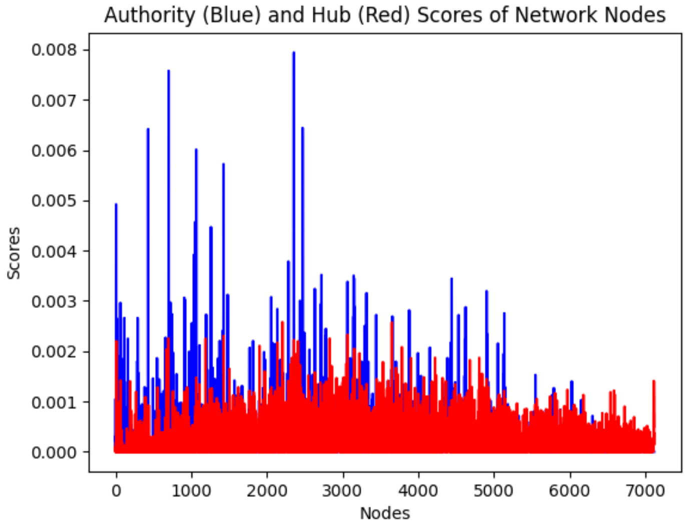
Q2.

Using the networkx library's pagerank() and hits() functions, we found the pagerank, authority and hub scores of all the nodes

The authority and hub scores of a graph are measures of importance assigned to nodes in the graph. The authority score of a node measures its importance based on the number and quality of incoming links to that node, while the hub score measures its importance based on the number and quality of outgoing links from that node.

From this we can infer that we have a small number of authority nodes (nodes with more incoming links/people receiving more votes) and the hub values (nodes with outgoing links/people a voting for others) are not too varying


PageRank Scores of Network Nodes


Authority (Blue) and Hub (Red) Scores of Network Nodes