Information Retrieval
Assignment-2

Akash Rawat (MT21005)
Shubham Rana (MT21092)

(Question 1)

For this question, we have used the following python libraries:
● Pandas, Numpy.
We have used the soc-sign-bitcoin-alpha: Bitcoin Alpha web of trust network dataset. The dataset has the following properties:

- Weighted
- Signed
- Directed
- Temporal

The number of nodes is 3,783 and the number of edges is 24,186.

We have now created the adjacency matrix and adjacency list for our dataset by iterating over the dataset and saving the properties of each node.

1. **Number of Nodes:** Using the adjacency list we have iterated over each node and saved it in a set to get the unique nodes.

```
The total number of nodes in the dataset is:  3783
```

2. **Number of Edges:** Using the adjacency matrix we have iterated over it and have counted the number of 1's in the matrix to get the number of edges.

```
The total number of edges in the dataset is:  24186
```

3. **Avg In-degree:** First we have calculated the in-Degree for each of the nodes and later on sum all the inDegree and divided by the total number of nodes.

```
The Average of Indegree of the nodes in the dataset is:  6.442727757059137
```

4. **Avg. Out-Degree:** First we have calculated the Out-Degree for each of the nodes and later on sum all the Out-Degree and divided by the total number of nodes.

```
The Average of Outdegree of the nodes in the dataset is:  7.360316494217894
```

5. **Node with Max In-degree:** We have already made a dictionary of the in-degree of each node and from there we have found the maximum indegree node.

```
The Maximum Indegree in the dataset is:  1
```

6. **Node with Max out-degree:** We have already made a dictionary of the out-degree of each node and from there we have found the maximum outdegree node.

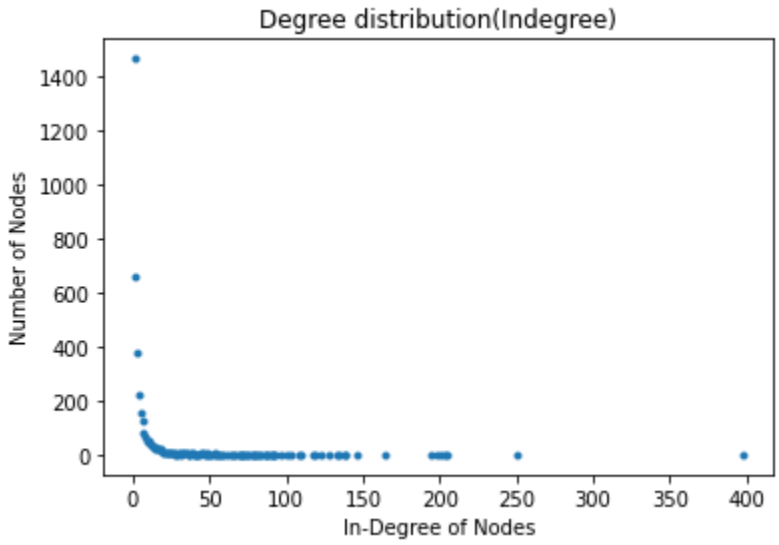
```
The Maximum Outdegree in the dataset is:  1
```

7. **The density of the network:** The density of the network is computed by the number of nodes divided by the total number of possible nodes in the network.

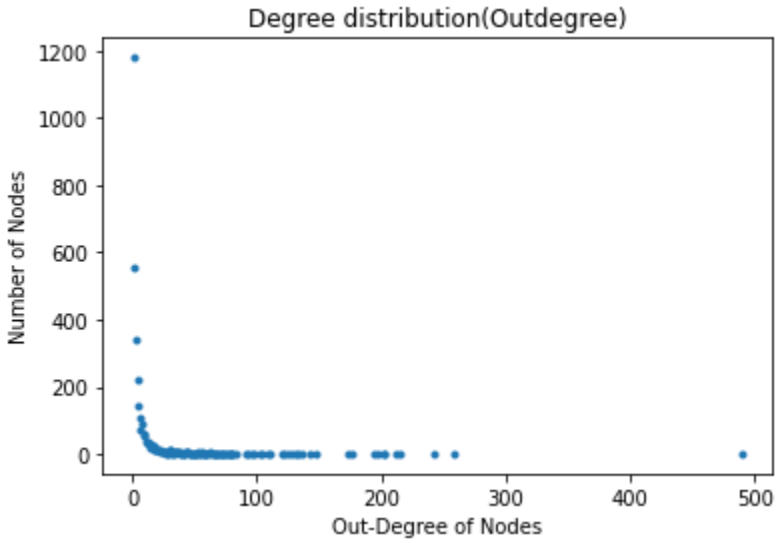
```
The density of the network of the dataset is:  0.0016904649973936393
```

Plot degree distribution of the network:

- **Indegree distribution of the network:**

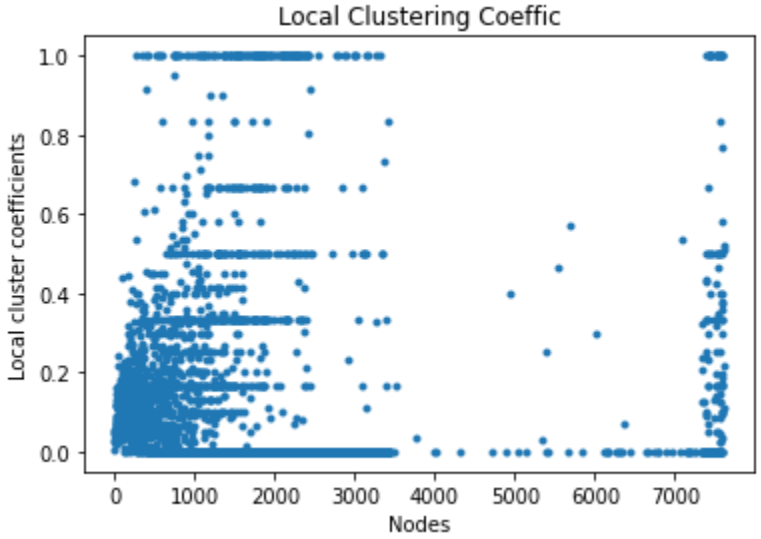


- **Outdegree distribution of the network:**



The local clustering coefficient of each node and plot of the clustering-coefficient distribution of the network:

To calculate the local clustering coefficient, we have gone through each of the nodes, and after that, we found all the neighbors of that node. After that, for all these neighbors we have seen how many links are there between each of its neighbor. Now we divide this value by the total possible number of links between these neighbors to get the local clustering coefficient for each node.



(Question 2)

1. **PageRank score for each node:**
   We have used the networkx library to obtain the page rank of the network.

```
{7188: 4.973996708553653e-05, 1: 0.016993099228405292, 430: 0.0002875150716955739,
```

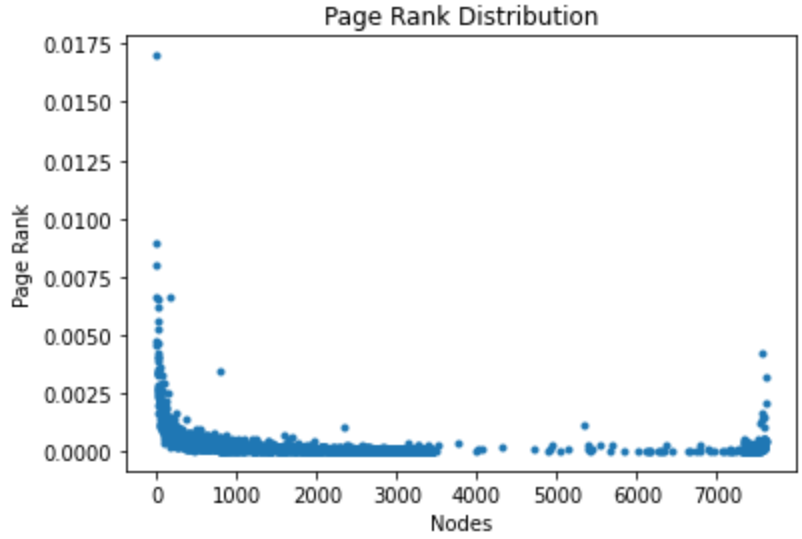
2. **Authority and Hub score for each node:**
   We have implemented authority and hub scores for our network from the scratch.
   Authority Score:

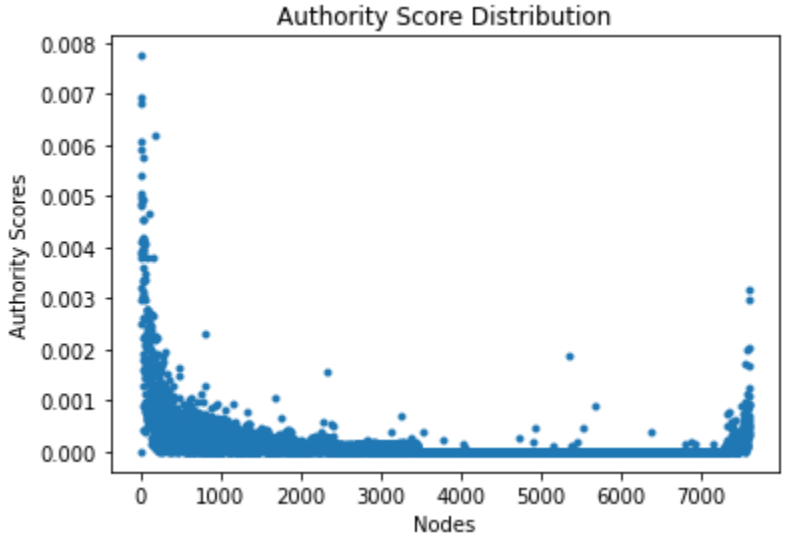
```
943: 0.00030404781421562217,
944: 0.00027164106510963295,
945: 0.00017029855451696586,
946: 0.000924570607229188,
947: 7.027334966040554e-05,
948: 4.780718261797479e-06,
949: 0.0006781856031482722,
950: 5.445762565456859e-06,
951: 0.0005415256393569937,
952: 5.378934421939674e-05,
953: 0.00020149560061563312,
954: 0.00012166011367830571,
955: 0.0004993684535297314,
956: 0.00023662845508485816,
```

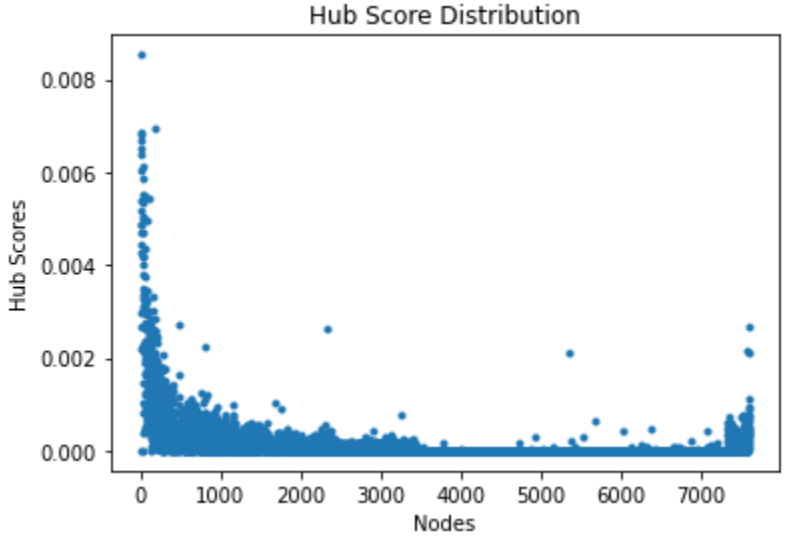
   Hub Score:

```
943: 0.0003616335252159231,
944: 0.00030634175866075413,
945: 0.00019108926964160893,
946: 0.0009629710617791111,
947: 6.97286809398037e-05,
948: 5.727777844023979e-06,
949: 0.0007067485385306671,
950: 5.851945406081773e-06,
951: 0.0006162094747895102,
952: 8.225135682500007e-05,
953: 0.00022121407842385247,
954: 0.0002284569443774923,
955: 0.0005649287517824817,
956: 0.0002562499819956003,
957: 5.9459384884932344e-05,
```

3. **Compare the results obtained from both the algorithms in parts 1 and 2 based on the node scores.**


Page Rank Distribution
Page Rank
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
1000
2000
3000
4000
5000
6000
7000
Nodes


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


Hub Score Distribution
Hub Scores
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
5000
6000
7000
Nodes

As we can see in the above graphs of different distributions such as page rank, hub-score, and authority score, the results are almost identical as hub and authority both are working on small subsets of the graph to find the score for each node, and page rank works on the probability distribution of all the nodes. So, we can conclude that all three of these techniques in this dataset give similar results.