ASSIGNMENT-3
INFORMATION RETRIEVAL

Team:
Yukti Goswami (MT21109)
Saurabh Pandey (MT21077)

Question 1 - [45 Points] Link Analysis

[2 points] Represent the network in terms of its 'adjacency matrix' as well as 'edge list'.

```
-------------- Adjacency Matrix -------------
      1     2     3     4     5     6     7     8     9     10    ...  5995  \
1        0     1     0     1     1     1     1     1     1     1  ...     0
2        1     0     0     1     0     1     1     0     0     1  ...     0
3        1     1     0     1     0     0     1     0     0     1  ...     0
4        1     1     0     0     0     1     1     0     0     0  ...     0
5        1     0     0     0     0     1     1     0     0     0  ...     0
...    ...   ...   ...   ...   ...   ...   ...   ...   ...   ...  ...   ...
6000     0     0     0     0     0     0     0     0     0     0  ...     0
6002     0     0     0     0     0     0     0     0     0     0  ...     0
6003     0     0     0     0     0     0     0     0     0     0  ...     0
6004     0     0     0     0     0     0     0     0     0     0  ...     0
6005     0     0     0     0     0     0     0     0     0     0  ...     0

      5996  5997  5998  5999  6000  6002  6003  6004  6005
1        0     0     0     0     0     0     0     0     0
2        0     0     0     0     0     0     0     0     0
3        0     0     0     0     0     0     0     0     0
4        0     0     0     0     0     0     0     0     0
5        0     0     0     0     0     0     0     0     0
...    ...   ...   ...   ...   ...   ...   ...   ...   ...
6000     0     0     0     0     0     0     0     0     0
6002     0     0     0     0     1     0     0     0     0
6003     0     0     0     0     0     0     0     0     0
6004     0     0     0     0     0     0     0     0     0
6005     0     0     0     0     0     0     0     0     0

[5881 rows x 5881 columns]
```

```
---------------- Edge List ---------------
[[6, 2], [6, 5], [1, 15], [4, 3], [13, 16], [13, 10], [7, 5], [2, 21], [2,
20], [21, 2], [21, 1], [21, 10], [21, 8], ........, [13, 1128], [1128,
13]]
```

[28 points] Briefly describe the dataset chosen and report the following:

1. Number of Nodes
2. Number of Edges
3. Avg In-degree
4. Avg. Out-Degree
5. Node with Max In-degree
6. Node with Max out-degree
7. The density of the network

About Dataset

ref: https://snap.stanford.edu/data/soc-sign-bitcoin-otc.html

This is a who-trusts-whom network of people who trade using Bitcoin on a platform called Bitcoin OTC. Since Bitcoin users are anonymous, there is a need to maintain a record of users' reputations to prevent transactions with fraudulent and risky users. Members of Bitcoin OTC rate other members in a scale of -10 (total distrust) to +10 (total trust) in steps of 1. This is the first explicit weighted signed directed network available for research.

Dataset statistics:

Nodes 5,881

Edges 35,592

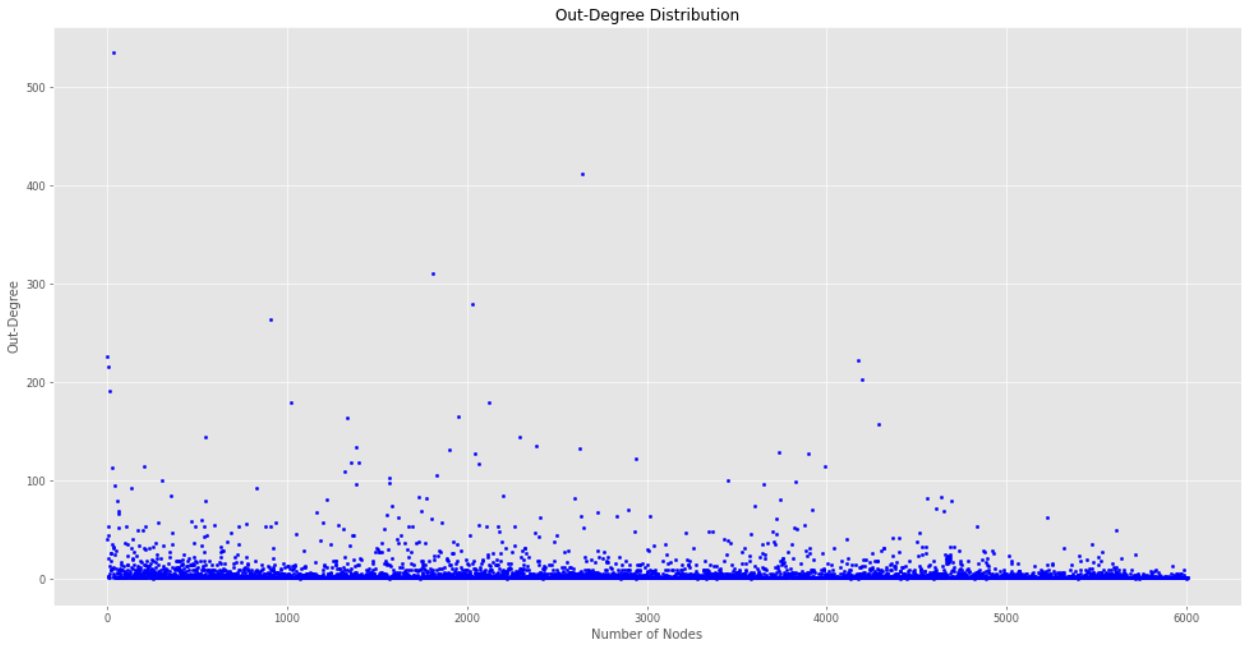
Range of edge weight -10 to +10

Percentage of positive edges 89%

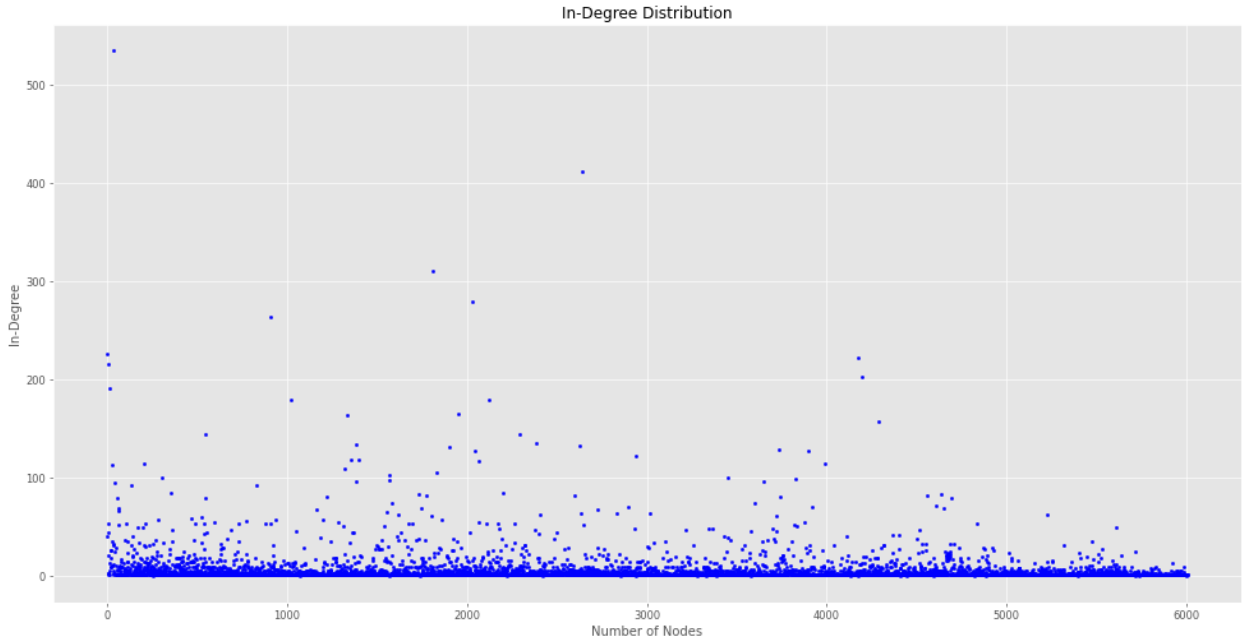
1. Number of Nodes: 5881
2. Number of Edges: 35592
3. Avg. In-degree: 6.052031967352491
4. Avg. Out-degree: 6.052031967352491
5. Node with Max in-degree: 35 in_degree: 535
6. Node with Max Out-degree: 35 out_degree: 535
7. The density of the network: 0.0010292571373048454

Further, perform the following tasks:
1. [5 points] Plot degree distribution of the network (in case of a directed graph, plot in-degree, and out-degree separately).

Out-Degree Distribution:



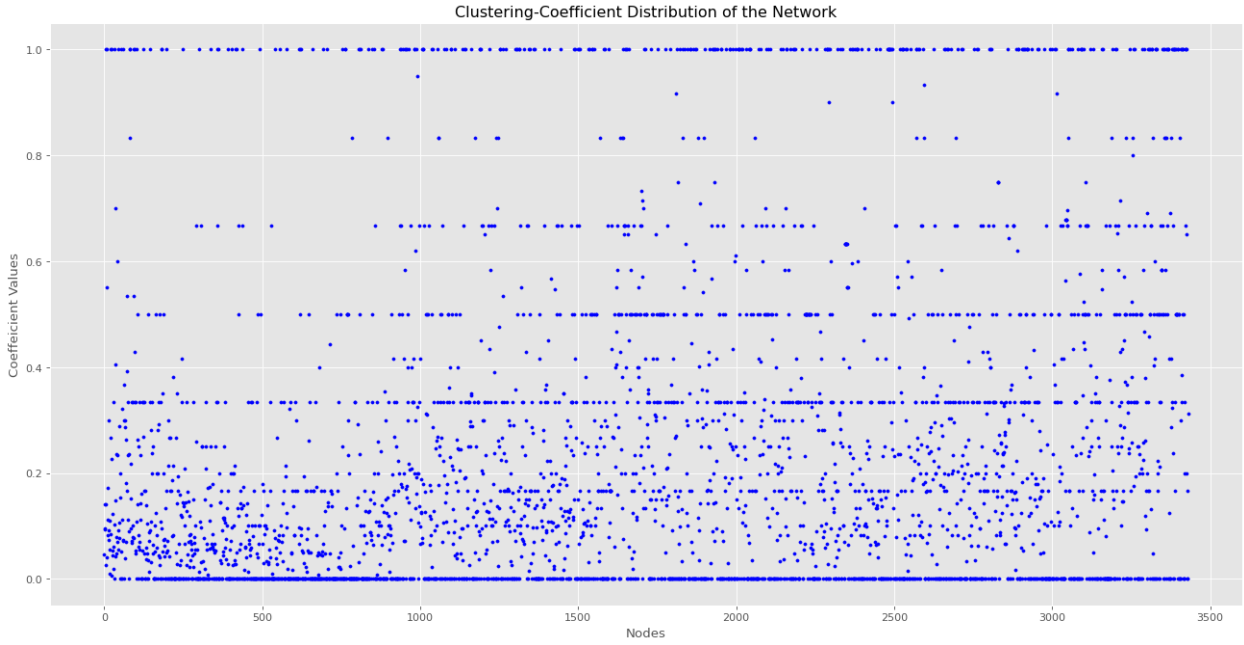
In-Degree Distribution:



2. [10 points] Calculate the local clustering coefficient of each node and plot the clustering-coefficient distribution of the network.

```
--------- Local Clustering Coefficient of Each Node ---------
[0.04576204523107178, 0.09451219512195122, 0.14047619047619048,
0.0936408106219427, 1.0, 0.1416490486257928, ……., 0.16666666666666666,
0.3111111111111111]

Clustering Coefficient Distribution
```



FORMULAS:

- Avg. In-Degree = Sum of number of incoming edges of every node / total number of nodes

- Avg. Out-Degree = Sum of number of outgoing edges of every node / total number of nodes
- Network Density (Directed Graph) = Number of edges / total number of nodes * (total number of nodes - 1)
- Local Clustering Coefficient of each node = Number of pairs of neighbors of the node that are connected / number of pairs of neighbors of the node

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

1. [15 points] PageRank score for each node

	Nodes	Page Rank Scores
0	6	0.000774
1	2	0.000977
2	5	0.000093
3	1	0.005029
4	15	0.000323
...	...	...
5876	6000	0.000035
5877	6002	0.000065
5878	6003	0.000047
5879	6004	0.000052
5880	6005	0.000052

5881 rows × 2 columns

2. [15 points] Authority and Hub score for each node

	Nodes	Authority Scores	Hub Scores
0	6	0.001572	0.001463
1	2	0.000589	0.000776
2	5	0.000170	0.000209
3	1	0.004496	0.004637
4	15	0.000295	0.000302
...	...	...	...
5876	6000	-0.000000	-0.000000
5877	6002	-0.000000	-0.000000
5878	6003	0.000002	-0.000000
5879	6004	0.000113	-0.000000
5880	6005	0.000113	-0.000000

5881 rows × 3 columns

[5 points] Compare the results obtained from both the algorithms in parts 1 and 2 based on the node scores.

	Nodes	Page Rank Scores	Authority Scores	Hub Scores
0	6	0.000774	0.001572	0.001463
1	2	0.000977	0.000589	0.000776
2	5	0.000093	0.000170	0.000209
3	1	0.005029	0.004496	0.004637
4	15	0.000323	0.000295	0.000302
...	...	...	...	...
5876	6000	0.000035	-0.000000	-0.000000
5877	6002	0.000065	-0.000000	-0.000000
5878	6003	0.000047	0.000002	-0.000000
5879	6004	0.000052	0.000113	-0.000000
5880	6005	0.000052	0.000113	-0.000000

5881 rows × 4 columns

Analysis:

Page Rank Scores

- PageRank computes a ranking of nodes in the graph based on the structure of the incoming links.
- Higher the incoming links, the higher the PageRank Score.
- Used for applying search after ranking all the nodes of the complete graph.

HITS

- HITS algorithm computes the authority score for a node based on the incoming links and computes the hub score based on outgoing links.
- HITS Score is higher if there are many good nodes as it is based on outgoing links.
- After completely searching the complete graph HITS is applied to the subgraph.