IR Assignment 3
Group 70

Submitted by:
Utsav Baghela (MT21101)
Ashwani Dongre (MT21016)

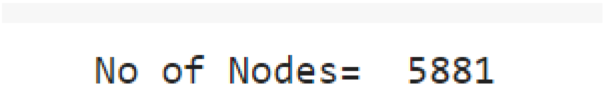
Dataset Link:
https://snap.stanford.edu/data/soc-sign-bitcoin-otc.html

Question 1 - Link Analysis

Adjacency Matrix: Firstly constructed a 2- Dimensional array of size n*n where n is the number of nodes. Each cell arr[i][j] contains 1 if there is a directed edge from node i to node j else the cell will contain 0.

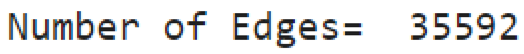
Edge List: This is a list of lists, where each list is of the format [i,j], denoting directed edge from ith node to jth node in the network.

1) **Number of Nodes:** Printed the number of nodes present in the dataset



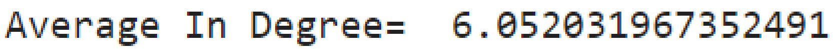
```
No of Nodes=  5881
```

2) **Number of Edges:** Printed the number of edges in the dataset. An edge from i to j is different from an edge from j to i as it is a directed graph.



```
Number of Edges=  35592
```

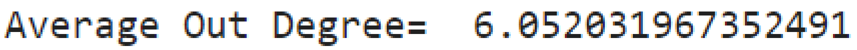
3) **Average In-degree:** The number of incoming edges on a vertex is the In-Degree of that vertex in a directed graph. The average of In-Degree is printed.
Avg In-degree = (sum of in-degrees of each node present) / number of nodes in dataset



```
Average In Degree=  6.052031967352491
```

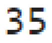
4) **Average Out-degree:** The number of outgoing edges of a vertex is the Out-Degree of that vertex in a directed graph. The average Out-Degree is printed.

   ○ Avg out-degree = (sum of out-degrees of each node present)/number of nodes present.



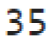
```
Average Out Degree=  6.052031967352491
```

5) **Node with Max In-degree:** Node with maximum In-Degree is printed.



```
35
```

6) **Node with Max out-degree:** Node with maximum Out-Degree is printed.

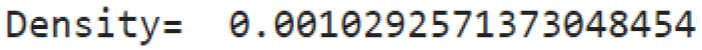


```
35
```

7) **The density of the network:** The density of a graph is the ratio of the number of edges present in the graph to the maximum number of edges that the graph can contain.

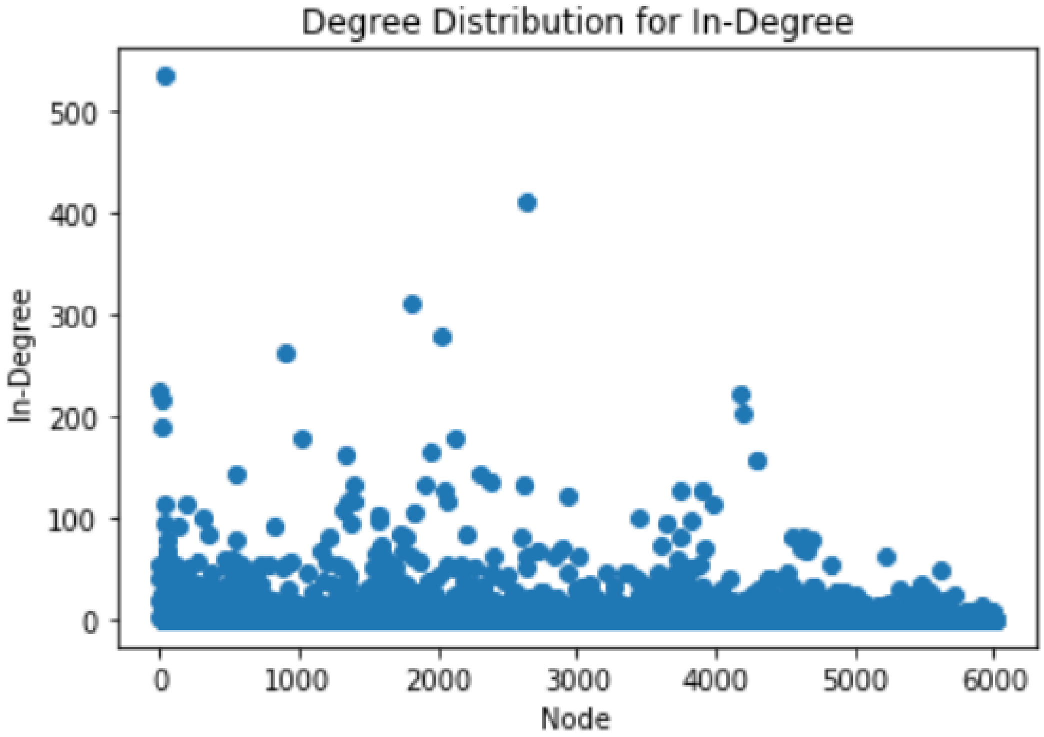
   ○ Density of the network = Actual Connections / Potential Connections

$$= \frac{len(Edge\ List)}{(number\ of\ nodes * (number\ of\ nodes - 1))}$$

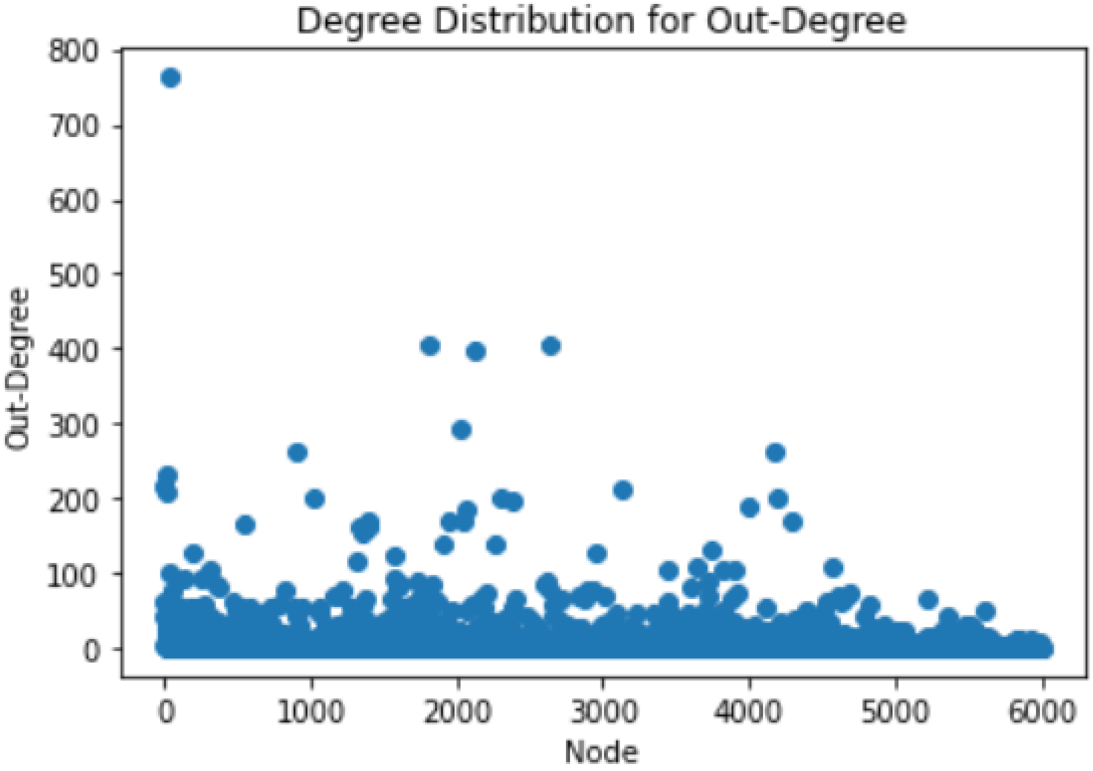


```
Density=  0.0010292571373048454
```

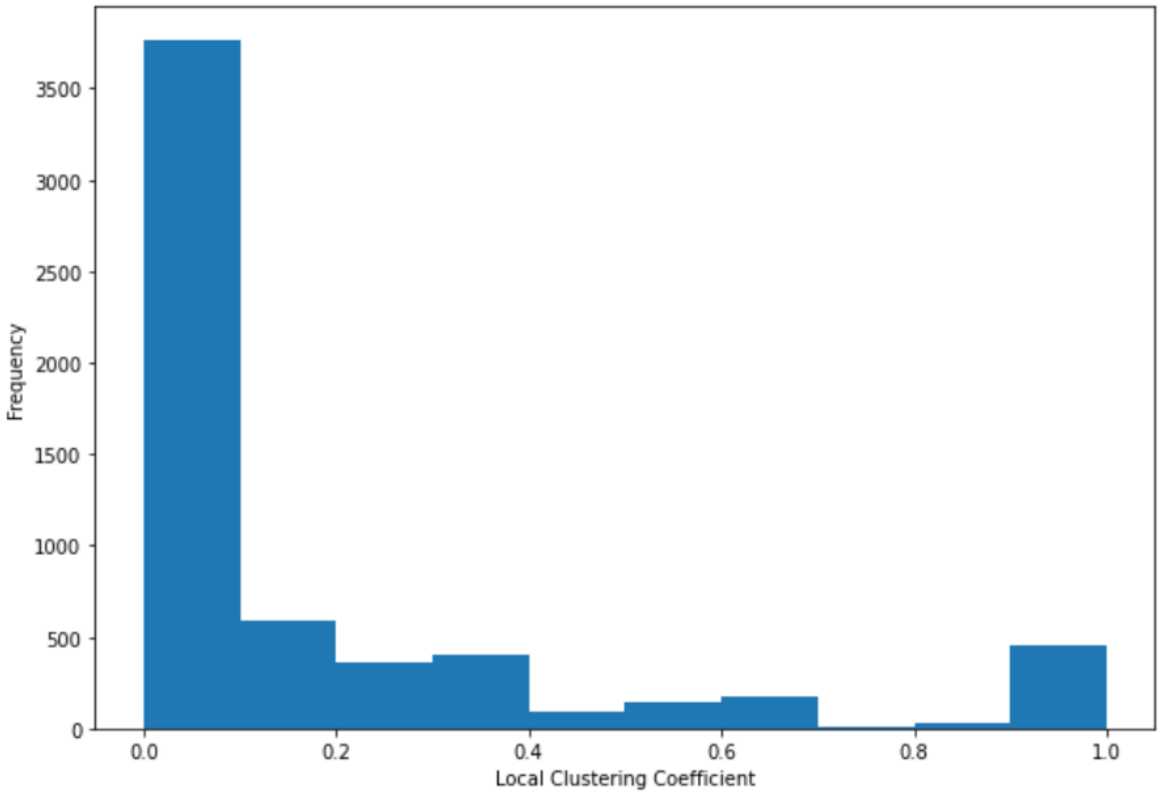
8) **Plot Degree Distribution (In-Degree):** Graph is plotted that is depicting the In-Degree corresponding to the nodes.



9) **Plot Degree Distribution (Out-Degree):** Graph is plotted that is depicting the Out-Degree corresponding to the nodes.



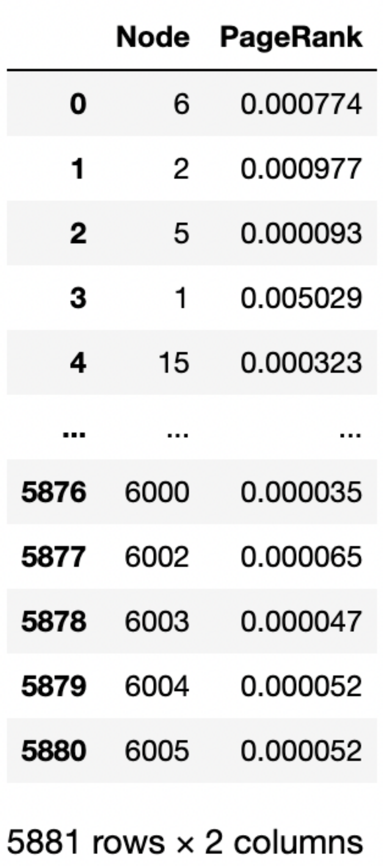
10) **Plot Local Clustering Coefficient: Local Clustering Coefficient** of a node is the proportion of the number of actual links between the neighbors of that node to the maximum possible links between those neighbors. Graph containing Local Clustering Coefficient and its frequency is plotted.


3500
3000
2500
2000
1500
1000
500
0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
Local Clustering Coefficient

Question 2

Page Rank:

PageRank computes a ranking of nodes in the graph based on the structure of the incoming links.
Approach : Di-Graph was created for each row of dataset, which contain source node and destination node and edge between that.
After creation of Di-Graph, networkx library was used to calculate the page rank for each node.



	Node	PageRank
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

Hubs Score:

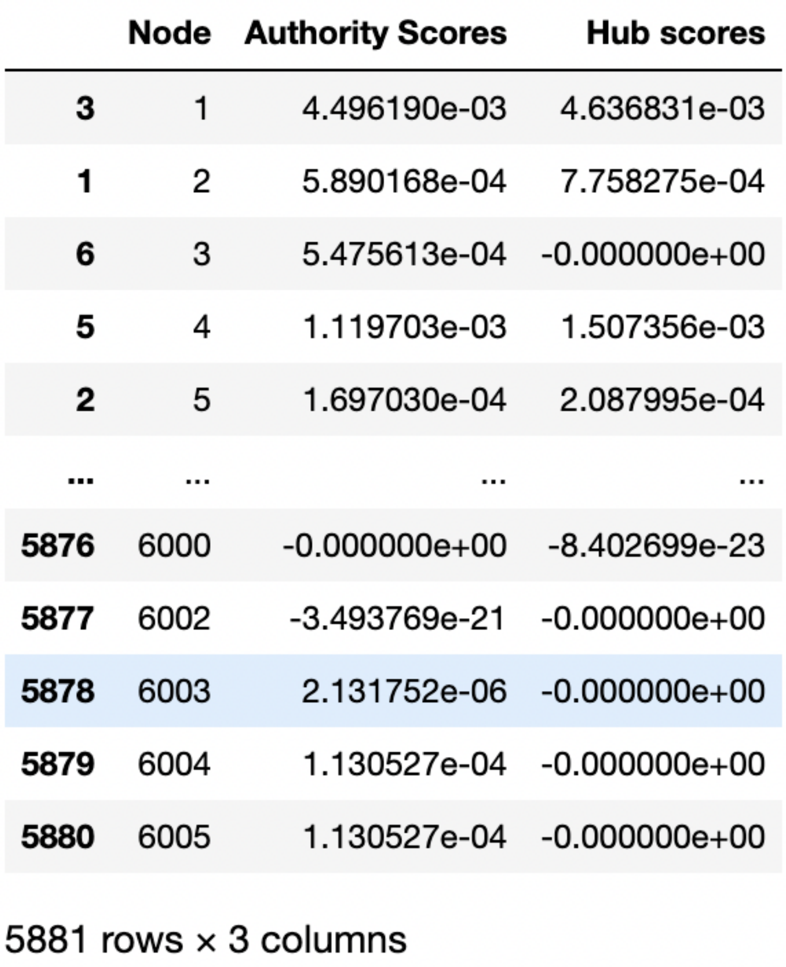
Hyperlink-Induced Topic Searchs a **link analysis algorithm** that rates Web pages,
It is also known as **hubs and authorities.**

Ref : https://en.wikipedia.org/wiki/HITS_algorithm

Authority Score:

Authority Score is our compound domain score that grades the overall quality of a website. If score is high the more assumed weight a domain's or webpage's will have.

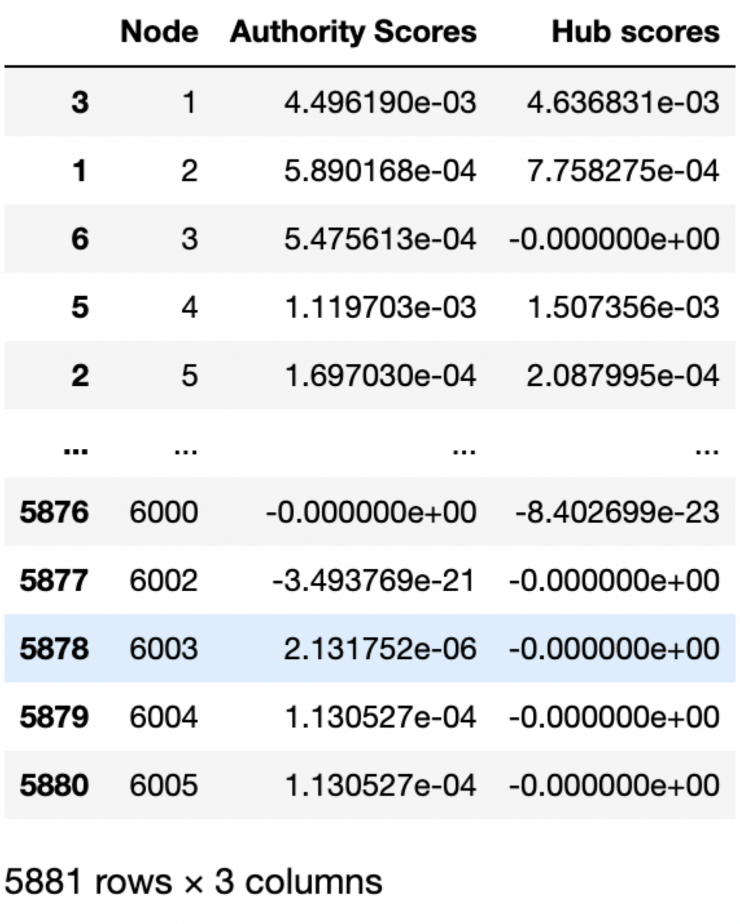
<u>Our Results</u> :



	Node	Authority Scores	Hub scores
3	1	4.496190e-03	4.636831e-03
1	2	5.890168e-04	7.758275e-04
6	3	5.475613e-04	-0.000000e+00
5	4	1.119703e-03	1.507356e-03
2	5	1.697030e-04	2.087995e-04
...	...	...	...
5876	6000	-0.000000e+00	-8.402699e-23
5877	6002	-3.493769e-21	-0.000000e+00
5878	6003	2.131752e-06	-0.000000e+00
5879	6004	1.130527e-04	-0.000000e+00
5880	6005	1.130527e-04	-0.000000e+00

5881 rows × 3 columns

Combined all scores values:

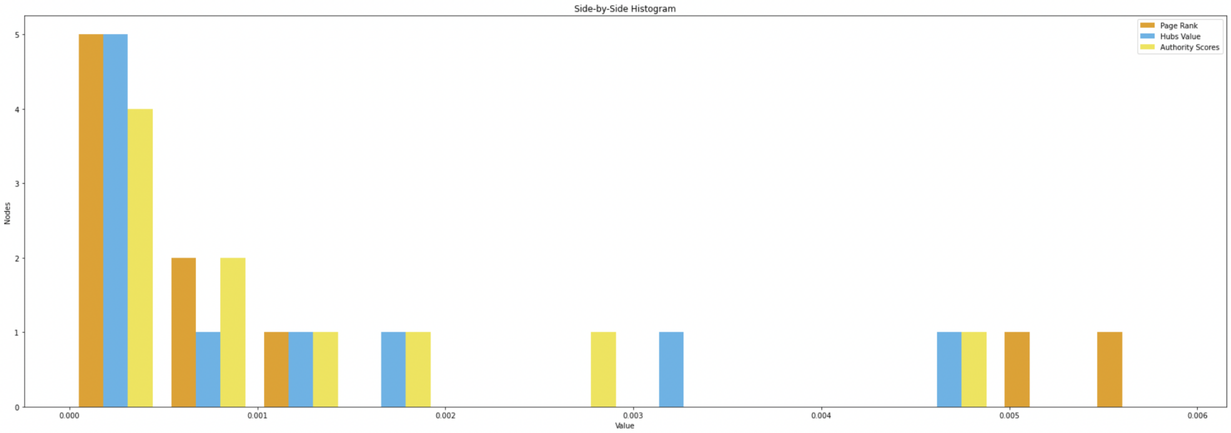


	Node	Authority Scores	Hub scores
3	1	4.496190e-03	4.636831e-03
1	2	5.890168e-04	7.758275e-04
6	3	5.475613e-04	-0.000000e+00
5	4	1.119703e-03	1.507356e-03
2	5	1.697030e-04	2.087995e-04
...	...	...	...
5876	6000	-0.000000e+00	-8.402699e-23
5877	6002	-3.493769e-21	-0.000000e+00
5878	6003	2.131752e-06	-0.000000e+00
5879	6004	1.130527e-04	-0.000000e+00
5880	6005	1.130527e-04	-0.000000e+00

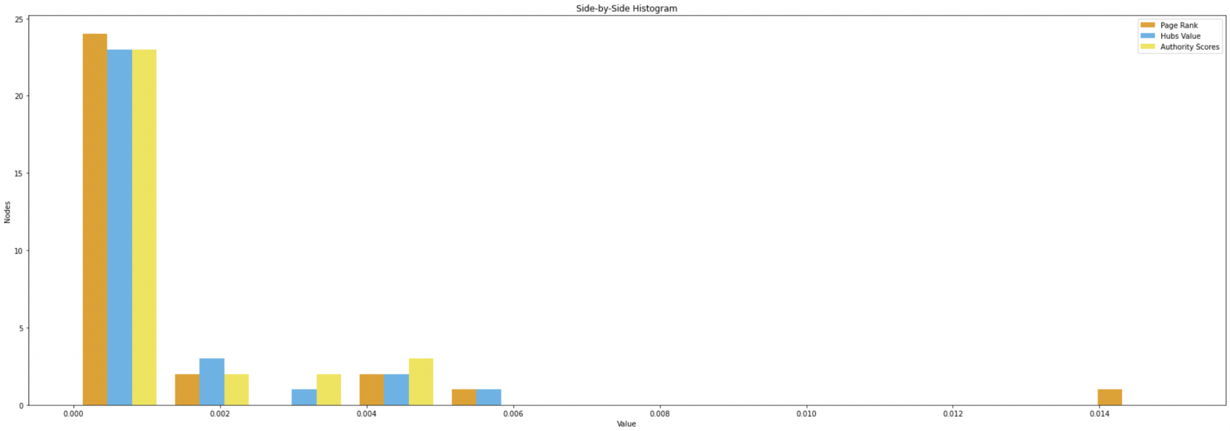
5881 rows × 3 columns

Comparing the results obtained from both the algorithms in parts 1 and 2 based on the node scores :

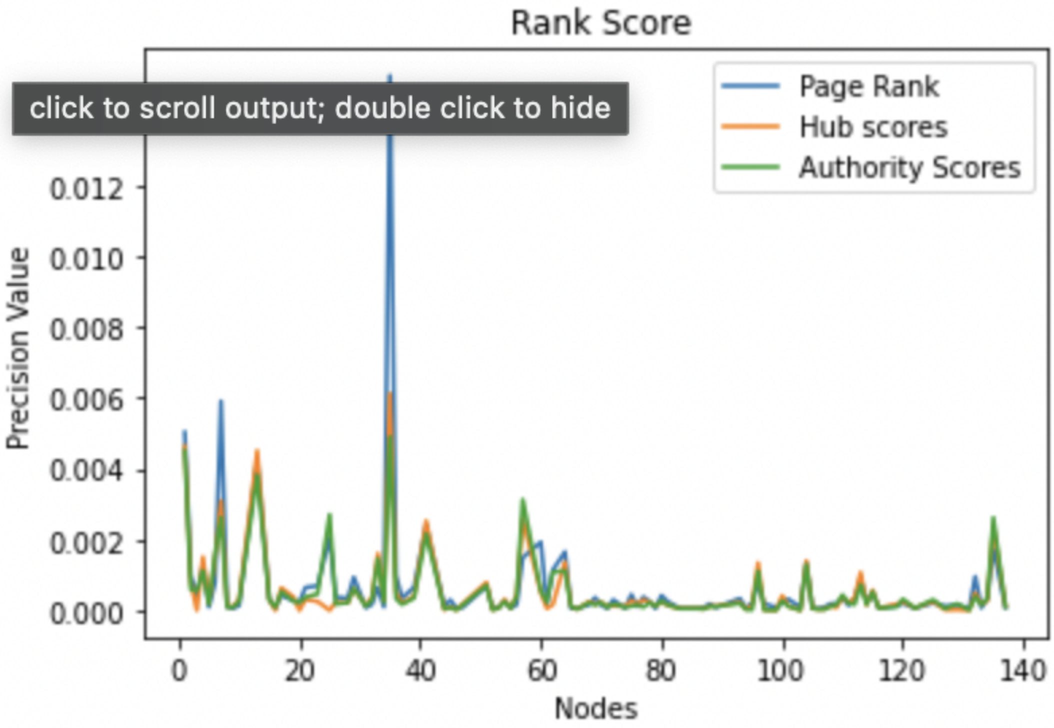
For First 10 node: (Side By side Historgram):



For First 30 node: (Side By side Historgram):



For all node: (Line Graph comparition):


Rank Score
click to scroll output; double click to hide
Page Rank
Hub scores
Authority Scores
Precision Value
Nodes