Assignment 6

Question 1:-

Preprocessing:- From the given choices I opted “Wiki-Vote.txt” dataset. I read the .txt file and then tokenize using the NLTK library and stored all the edge relations in a list.

Methodology:- I used above-preprocessed edge list and make 2 separate list which is Out degree nodes and corresponding indegree nodes and then I used following formulas to calculate different quantities as following:-

1.Number of Nodes: -
n=length of node list

result:- n=NO of Total Nodes: 7115

2. Number of Edges:-
e=lenght of all outgoing nodes list

result e= No of Edges: 103689

3. Avg In-degree:-
avgInDegree= e/n

result:- avgInDegree=14.573295853829936

4. Avg. Out-Degree
avgOutDegree= e/n

result:- avgOutDegree=14.573295853829936

5. Node with Max In-degree:-
maxInDegreeNode=mode(list_nodes_to)
=4037
maxInDegree=list_nodes_to.count(maxInDegreeNode)
=457

6. Node with Max out-degree
maxOutDegreeNode=mode(list_nodes_from)
=2565
maxOutDegree=list_nodes_from.count(maxOutDegreeNode)
=893

Note:- Both of the above can be calculated using the adjacency matrix. But the above solution is way faster than adjacency matrix. For Alternate solution, I also calculated the same using the adjacency matrix.

7. Density of the network:
dencity=e/possible_no_of_edges

possible_no_of_edges=n*(n-1)/2

Result:- dencity=0.004097075022161917

Ref:- https://www.the-vital-edge.com/what-is-network-density/

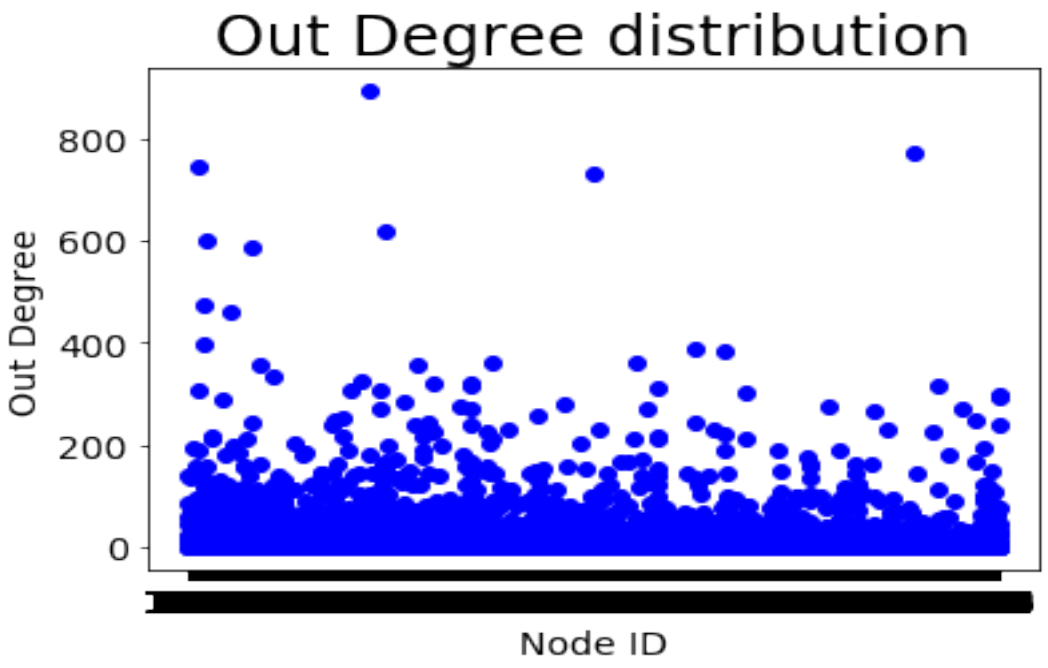
Other Tasks:-

1. Plot degree distribution of the network:-I plotted all the degrees according to their **lexicographically sorted node Ids**.

Plots:-

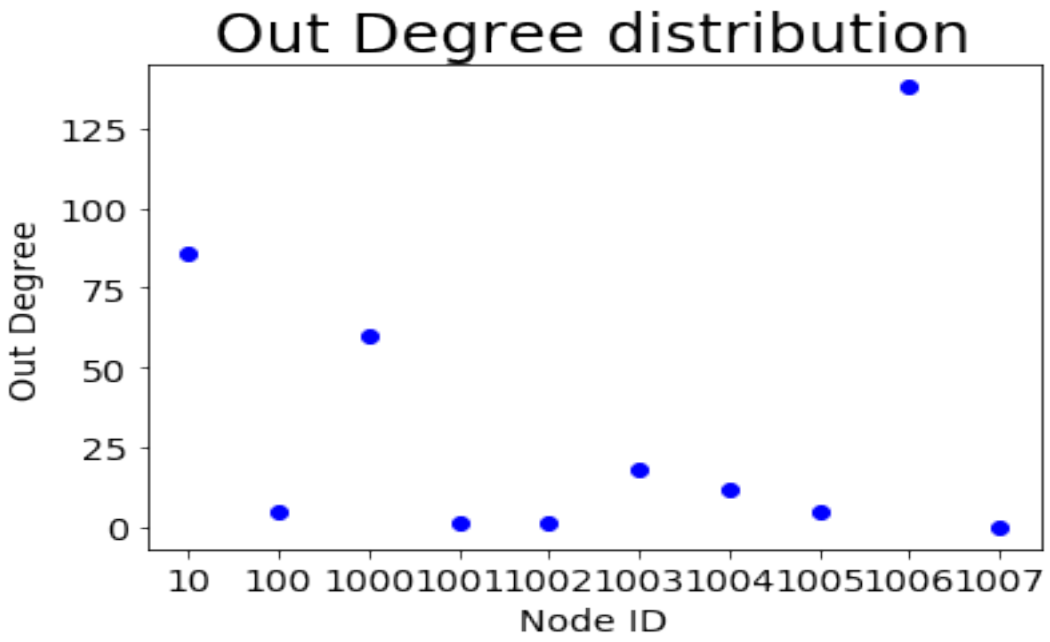
Note:- Because of 7115 nodes Node Id is not visible in the graph so I am also plotting same graph with less number of Node Ids as following:-

Out Degree distribution(At all the nodes):-

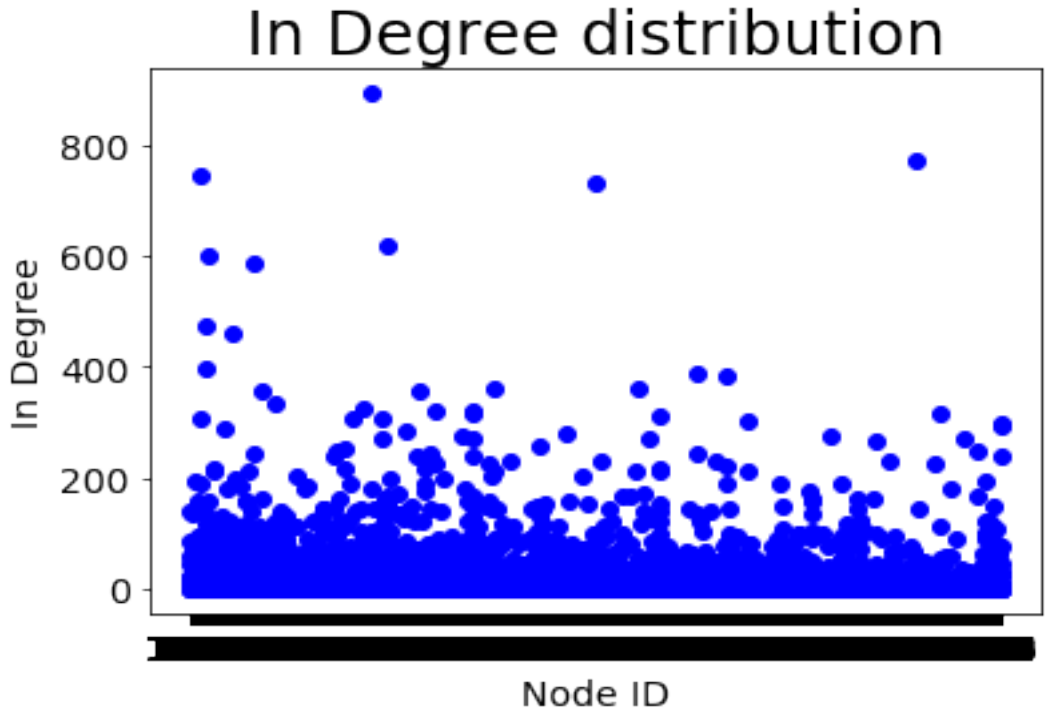


Out Degree Distribution at 10 nodes:-

This plot is for showing node ids with respect to the above graph.

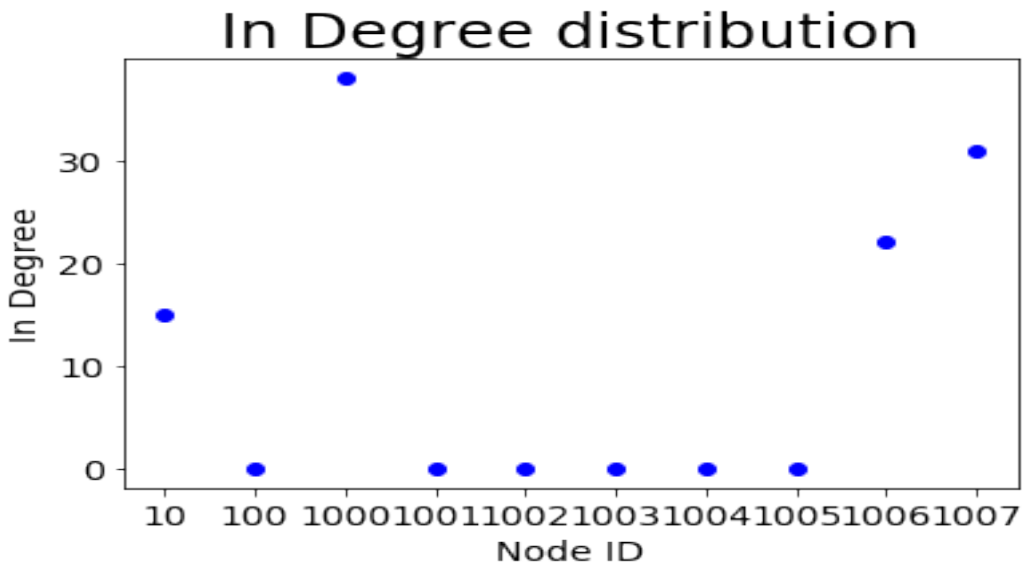


In Degree Distribution:-



Out Degree Distribution at 10 nodes:-

This plot is for showing node ids with respect to the above graph.



2. The clustering coefficient of each node:

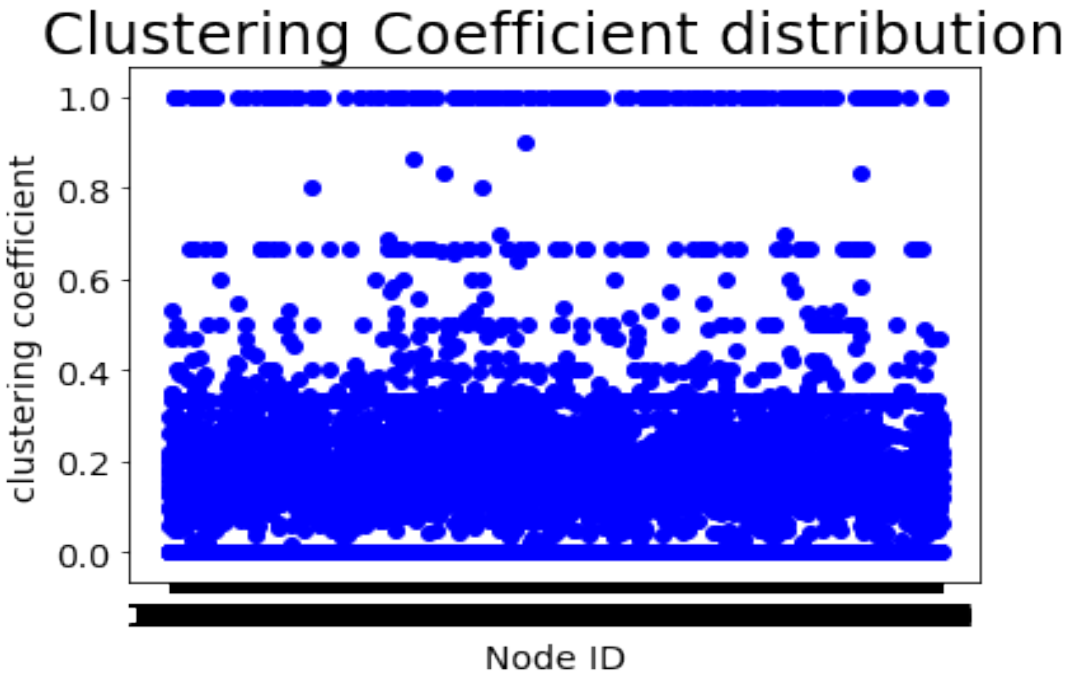
Ref:-https://en.wikipedia.org/wiki/Clustering_coefficient

As given in the question that we have to calculate the clustering coefficient of all the nodes it leads to local clustering coefficient.

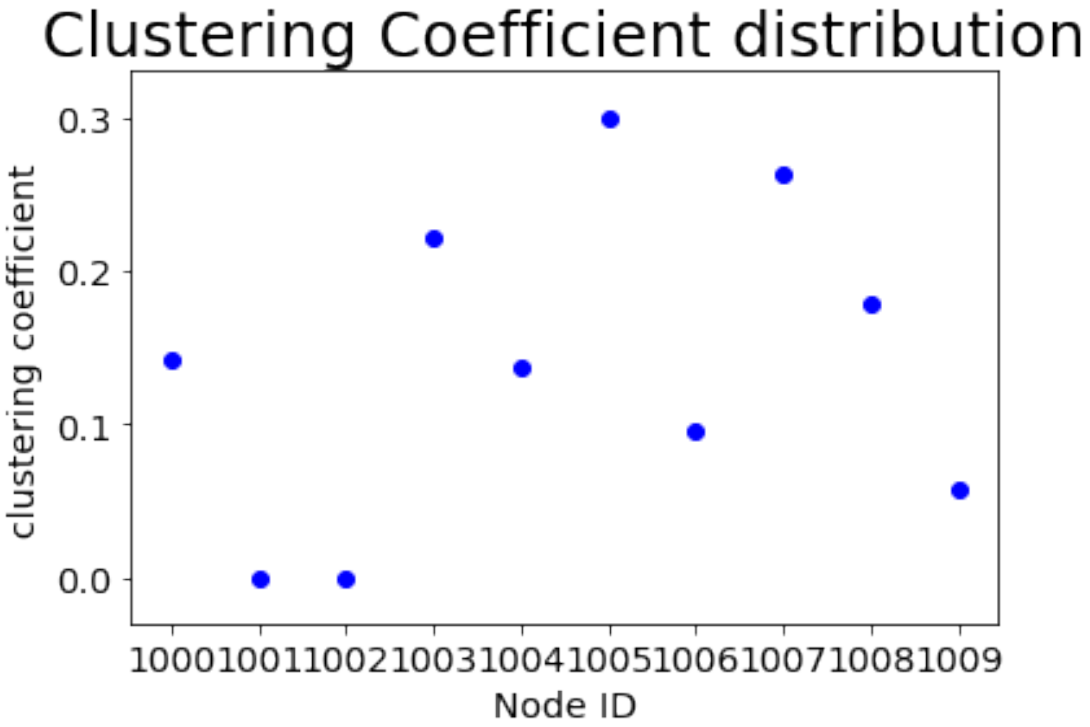
C_i=(edges in the cluster with nearest neighbor of node i)/(k*(k-1))
k= No of neignbor nodes with node i

Average clustering coefficient = 0.14

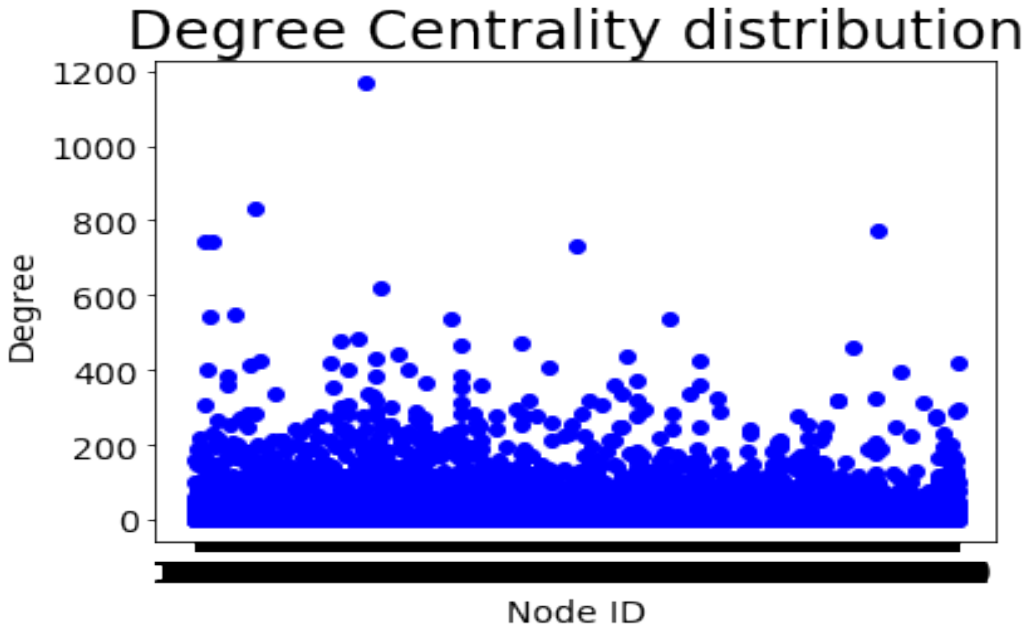
Note:- Here out-degree nodes and in-degree nodes, both are being considered as neighbor node to the i node.



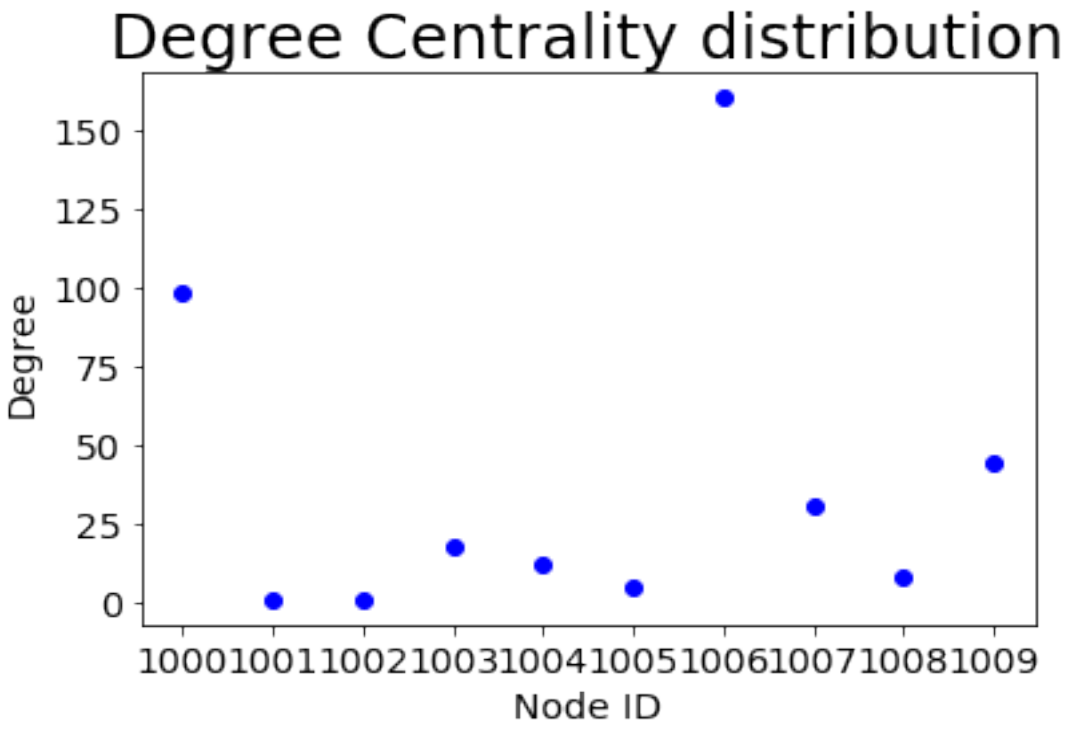
The same plot with less nodes:-



3. Degree centrality measure for each node:- For this centrality measure, I sum both indegree and outdegree of each node.
Total degree=indegree+outdegree
Degree centrality plot:- Here node ids are in lexicographical order.



Note:- In the above graph node ids are not visible because of its huge number so I am plotting the same graph with less no of node ids as following:-



Question 2:-

All the values of PageRank, hub, and authorities scores obtained from Networkx library are mentioned in the jupyter notebook output.

For each node, there is comparison between all the above 3 entities with respect to In Degree and Out Degree of that particular node.

Node Id observation based on sorted page rank:-

Node Id	In Degree	Out Degree	Page Rank	Hub	Authority
4037	457	15	0.0046127158911675485	0.00018275048104661505	0.002573241124142803
15	361	50	0.0036812207295292792	0.00041573260339102593	0.002201543492543811
6634	203	3	0.003524813657640259	2.2090351939381948e-05	0.0011684161446831453
2625	331	0	0.0032863743692309023	0.0	0.0021978968035237852
2398	340	62	0.0026053331717250192	0.0008125149592320502	0.002580147178008918
2470	149	0	0.0025301053283849546	0.0	0.0007495513993753806
2237	181	241	0.002504703800483994	0.003075961697014788	0.0012519304719652664
4191	259	20	0.0022662633042363454	0.00027639103834611	0.0020811941305208825
7553	190	0	0.002170185049195958	0.0	0.0011865887766775688
5254	265	33	0.0021500675059293235	0.0003772096308658232	0.0018247396642980143
1186	193	0	0.0020438936876029153	0.0	0.001000980405726099
2328	266	215	0.0020416288860889186	0.002322921611277255	0.0021723715454129585
1297	309	76	0.0019518608216122285	0.0008523342729841223	0.00225014463679536
4335	228	32	0.0019353014475784877	0.0003915264768396083	0.0018627007058849888
7620	208	0	0.0019301193957548752	0.0	0.001217113645661963
5412	219	0	0.00191670807752399	0.0	0.0018694113161459407
7632	178	0	0.0019037739909136618	0.0	0.001143052531112264
4875	181	0	0.0018675748225119092	0.0	0.0014027305777936597
3352	264	273	0.0017851250122027215	0.0033814231063793564	0.002328415091537902
2654	213	104	0.0017693207143482425	0.0011395717077516726	0.0017237094176338033
6832	170	0	0.0017646895191923734	0.0	0.0012335063414913119
762	272	50	0.0017478626294191988	0.0004017003019819967	0.00225587485637424
6946	117	68	0.0017404328450373553	0.0002973159824737583	0.000935775612563772
737	231	232	0.0017365555312247153	0.0016891026504129677	0.002039382629304915
2066	254	0	0.0017190133175865356	0.0	0.0021070409397065445
8293	164	0	0.0017044691035007064	0.0	0.001408707080985595
3089	244	4	0.0016993720805897224	6.85956207282843e-05	0.0022534066884266454
28	122	133	0.0016986730322136937	0.0003569336095454782	0.00025475061392999266
2535	232	47	0.0016682067978015206	0.0006935231001869295	0.0018680659041328002
214	175	0	0.001659919966936546	0.0	0.0008251269692542302
3334	217	13	0.0016535779185613182	0.0001900937838044287 3	0.0017510881860138216
665	97	7	0.0016523142774243047	7.717786278205172e-05	0.000387126946534206
4735	195	0	0.001620956758041132	0.0	0.0015184596329721948
6774	113	24	0.001615901009929992	0.00023652014561293498	0.0009003614613541905
7092	155	93	0.0015973872534344033	0.0007491594836652265	0.0011309196820779776
2565	274	893	0.0015503503379620108	0.00794049270807403	0.002223564103945871
5484	153	22	0.0015430456206335557	0.0003241146424281476	0.001279818744477702
4310	131	278	0.0014598839350747038	0.0028146321646042856	0.0011938378722827164
1211	188	79	0.001417714690132005	0.0007037165613430748	0.00134342468100998
5423	97	48	0.0014139627646056567	0.0004202980440707628	0.0010337733620430333

Here we can get some following comparison points:-

1. Page rank is based on in-degree of that node and authority score of the nodes from where it is being referred (from the nodes from where we have incoming links).
2. In ideal case for high page-rank node should be a sink node.
3. Hub score is based on out-degree of a node.
4. In ideal case, hub should be a source node.
5. Authority score is based on incoming links or indegree of a node.