IR ASSIGNMENT 6 ANALYSIS FILE

The graph I have used for both the questions is the Wikipedia Vote network, the link for which is given below:

https://snap.stanford.edu/data/wiki-Vote.html

Question 1: The description of the dataset used is:

A small part of Wikipedia contributors are administrators, who are users with access to additional technical features that aid in maintenance. In order for a user to become an administrator a Request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote decides who to promote to adminship.

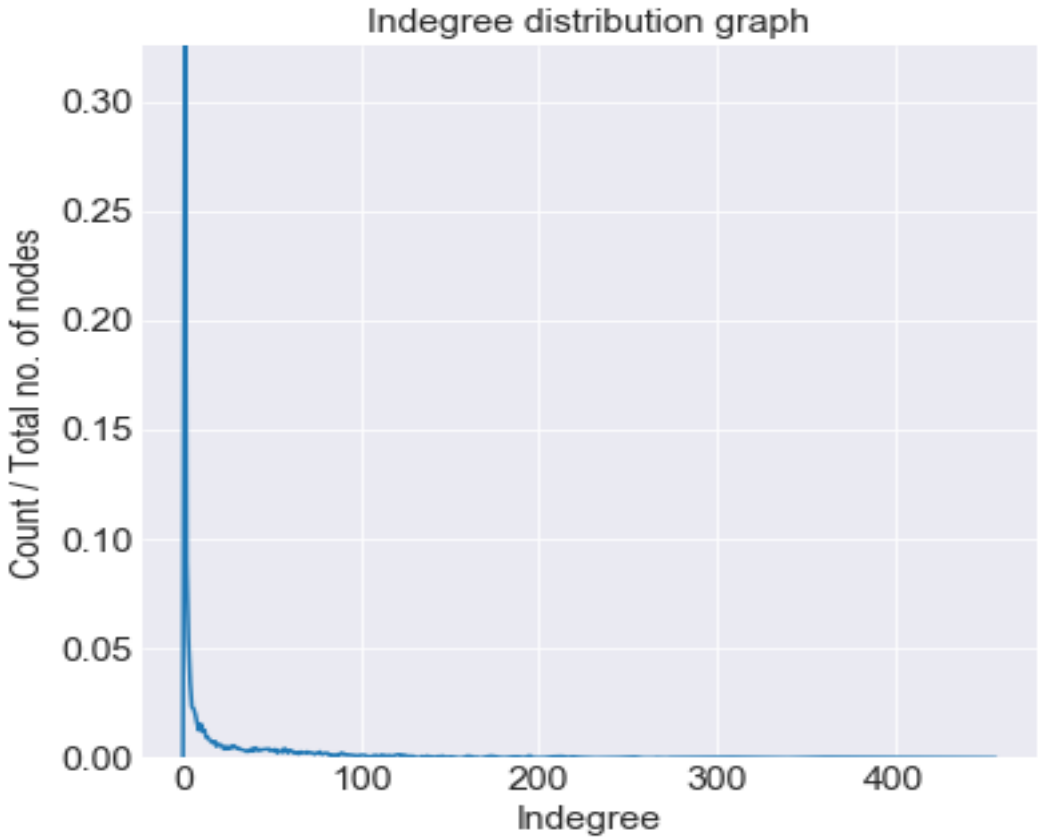
The network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008. Nodes in the network represent wikipedia users and a directed edge from node *i* to node *j* represents that user *i* voted on user *j*. Thus this is a directed graph/

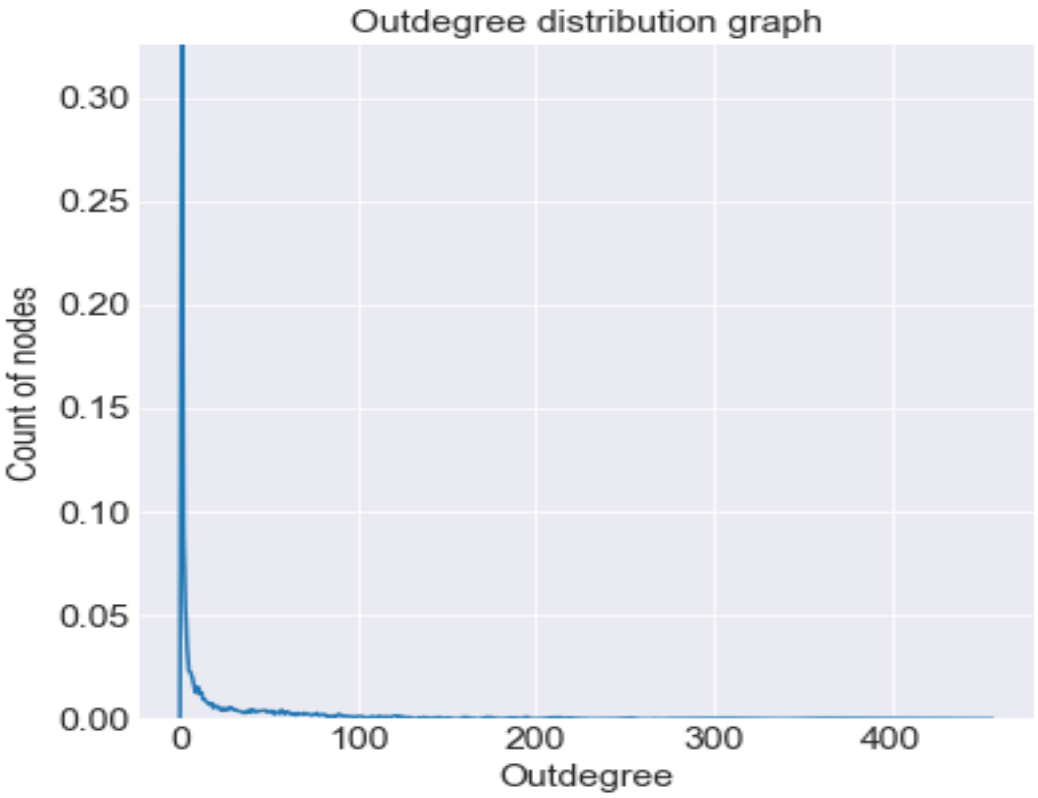
The different values for the graph are:

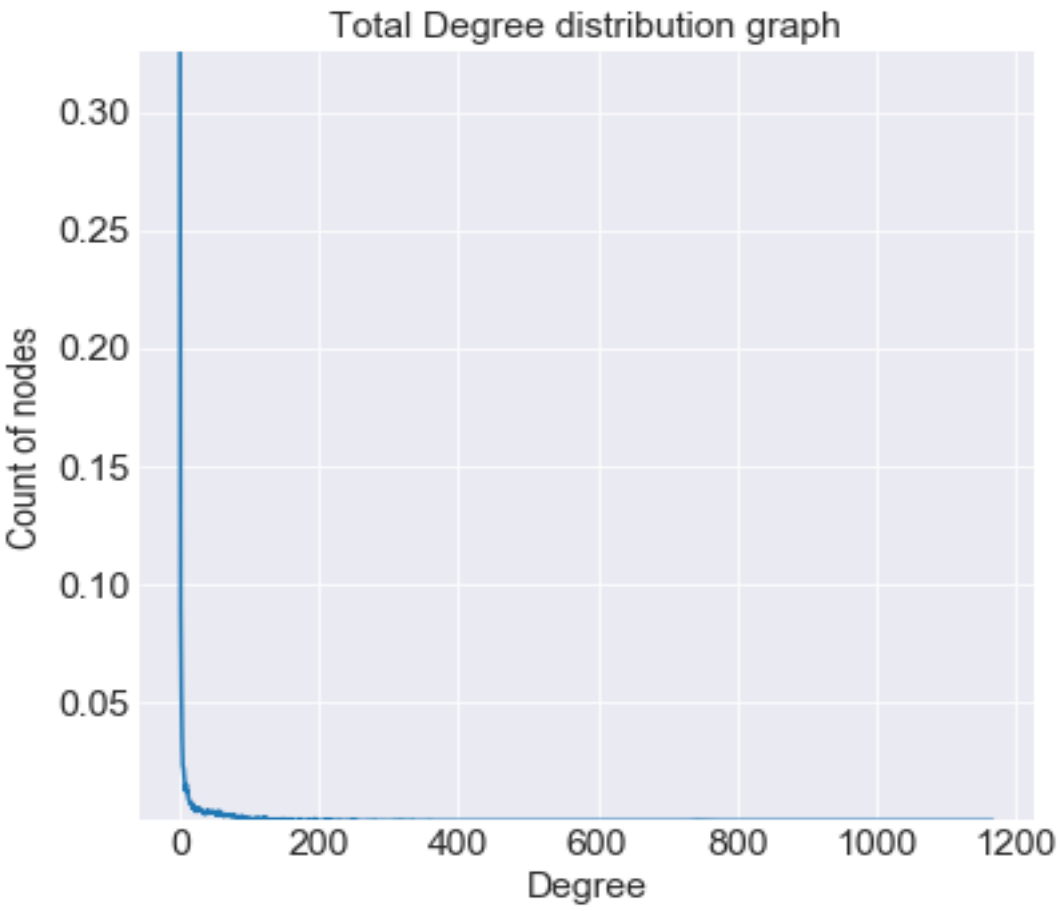
1. The no. of nodes in Wiki-Vote graph = 7115
2. The no. of edges in Wiki-Vote graph = 103689
3. The average indegree for wiki-vote graph = 14.573295853829936
4. The average outdegree for Wiki-Vote graph is = 14.573295853829936
5. The node with maximum indegree is = 4037
6. The node with maximum outdegree is = 2565
7. The formula used for calculating the density of the network is:
   Density = Total no. of edges / (Total no. of nodes * (Total no. of nodes - 1))
   The density of the Wiki-Vote network is = 0.0020485375110809584

The degree distribution and other values are:

1. For finding the degree distributions I have taken the degree on X- axis and the value (no. of nodes with that degree / Total no. of nodes) on Y-axis. Thus for a particular degree value on the X-axis we have on Y-axis the probability value of a randomly chosen node having that particular degree. The degree distribution plots are:




Outdegree distribution graph
Count of nodes
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
100
200
300
400
Outdegree


Total Degree distribution graph
Count of nodes
0.30
0.25
0.20
0.15
0.10
0.05
0
200
400
600
800
1000
1200
Degree

2. The clustering coefficient of the nodes:
   The local clustering coefficient is calculated for each of the nodes using the formula given below:

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

   The above formula calculated the local clustering coefficient of node i and N_i is the set of neighbours of node i and k_i is the number of neighbours of node i where the definition of neighbours is:

$$N_i = \{v_j : e_{ij} \in E \vee e_{ji} \in E\}.$$

   i.e. all the nodes to which node 'i' has out links as well as the nodes from which node I has in links.

   The average clustering coefficient value is found by taking the average of the local clustering coefficient values of all the nodes.

   The average clustering coefficient of Wiki-Vote network = 0.08053412949449601

   The top 10 nodes according to clustering coefficient value are:
   1. 5310 1.0
   2. 7029 1.0
   3. 2293 1.0
   4. 1782 1.0
   5. 7383 1.0
   6. 4799 1.0
   7. 4922 1.0
   8. 1923 1.0
   9. 7842 1.0
   10. 7302 1.0

   The local clustering coefficient value of all nodes is printed in the ipynb file.

3. The centrality value:
   Here I have found the degree centrality value of each node. Formula used is:

   Degree centrality value = Indegree value + Outdegree value

   The top 10 nodes according to degree centrality values along with the values are:

   NodeId Degree centrality value
   1. 2565 1167
   2. 1549 832
   3. 766 773

4. 11 743
5. 1166 743
6. 457 732
7. 2688 618
8. 1374 551
9. 1151 543
10. 5524 538

The degree centrality values of all nodes are given in the ipynb file.

Question 2: The WikiVote graph is loaded as a network directed graph and the different scores are calculated for each node:

The top 10 nodes according to the PageRank score along with their PageRank score are:

	Node id	PageRank score
1.)	4037	0.0046127158911675485
2.)	15	0.0036812207295292792
3.)	6634	0.003524813657640259
4.)	2625	0.0032863743692309023
5.)	2398	0.0026053331717250192
6.)	2470	0.0025301053283849546
7.)	2237	0.002504703800483994
8.)	4191	0.0022662633042363454
9.)	7553	0.002170185049195958
10.)	5254	0.0021500675059293235

The top 10 nodes according to the hub score along with their hub score are:

	Node id	Hub score
1.)	2565	0.00794049270807403
2.)	766	0.007574335297444512
3.)	2688	0.006440248991012525
4.)	457	0.00641687049019565
5.)	1166	0.006010567902433343
6.)	1549	0.0057207540583986485
7.)	11	0.004921182064008282
8.)	1151	0.004572040701802756

9.) 1374 0.004467888792672376

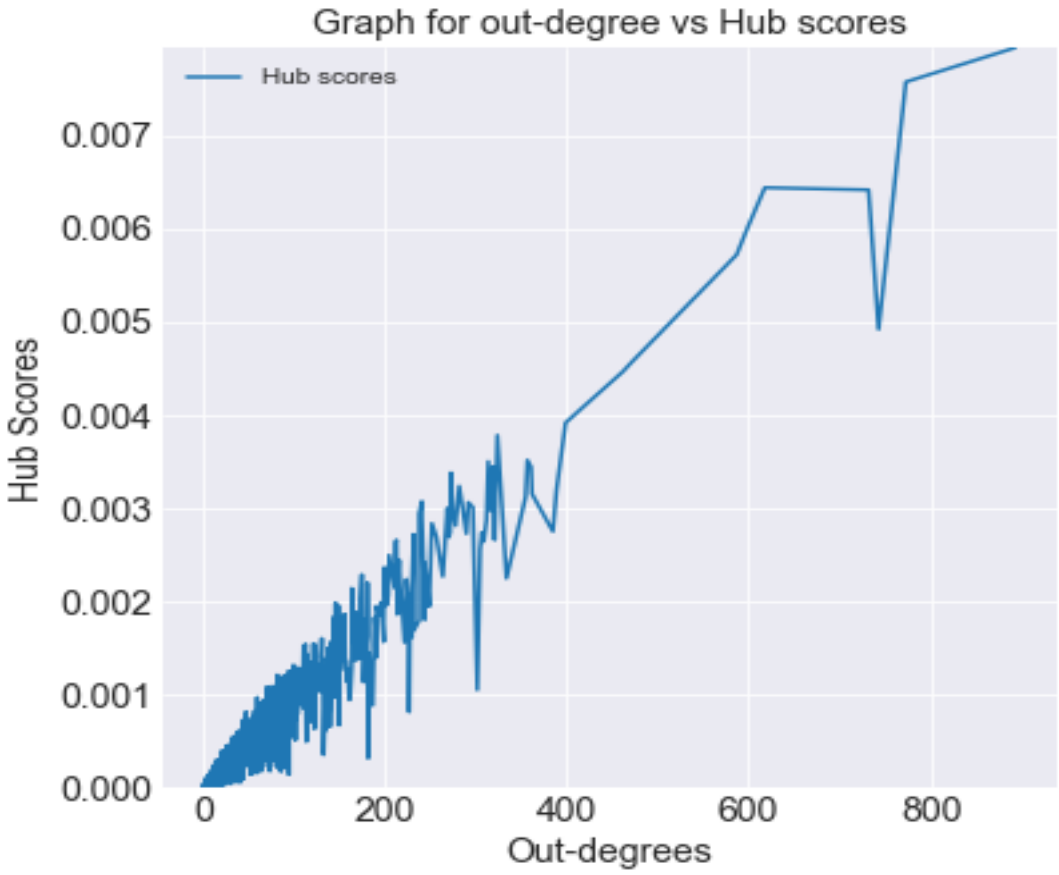
10.) 1133 0.003918881732047633

The top 10 nodes according to the Authority score along with their Authority score are:

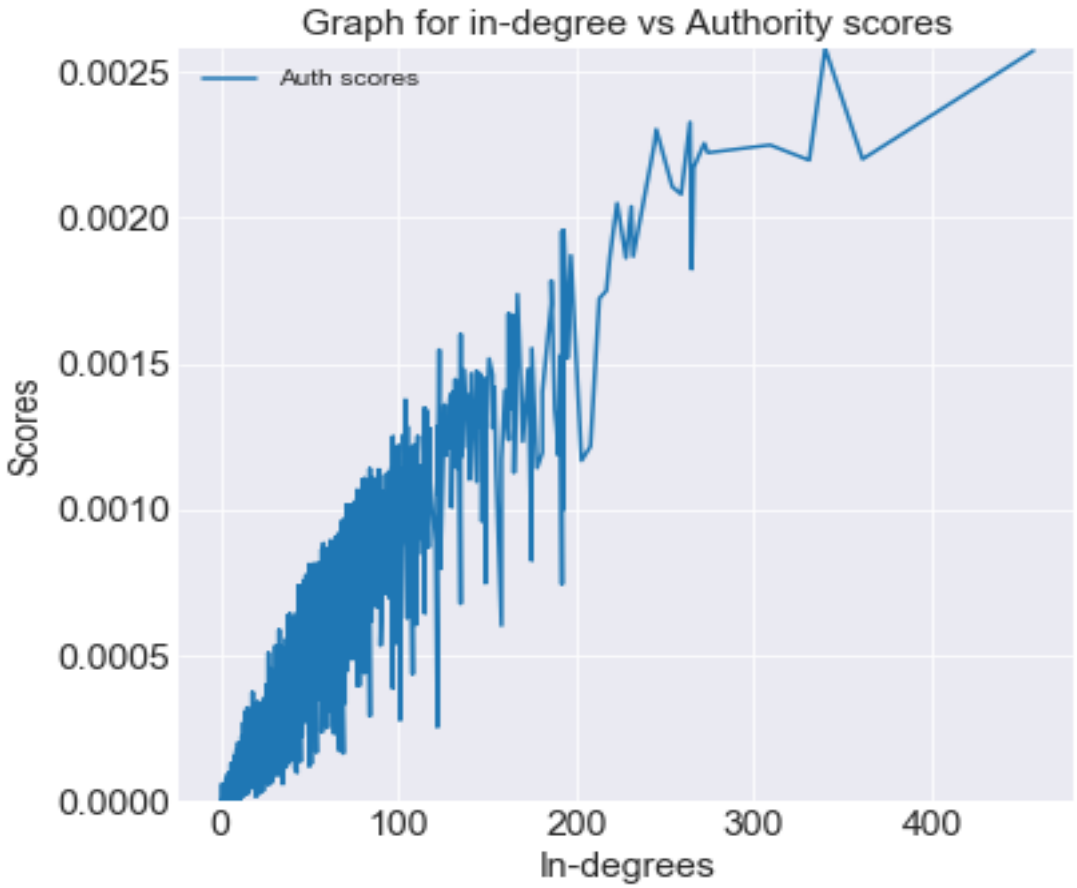
	Node id	Authority score
1.)	2398	0.002580147178008918
2.)	4037	0.002573241124142803
3.)	3352	0.002328415091537902
4.)	1549	0.0023037314804751075
5.)	762	0.00225587485637424
6.)	3089	0.0022534066884266454
7.)	1297	0.00225014463679536
8.)	2565	0.002223564103945871
9.)	15	0.002201543492543811
10.)	2625	0.0021978968035237852

The PageRank values, Hub scores and Authority scores for all nodes are printed in the ipynb files.

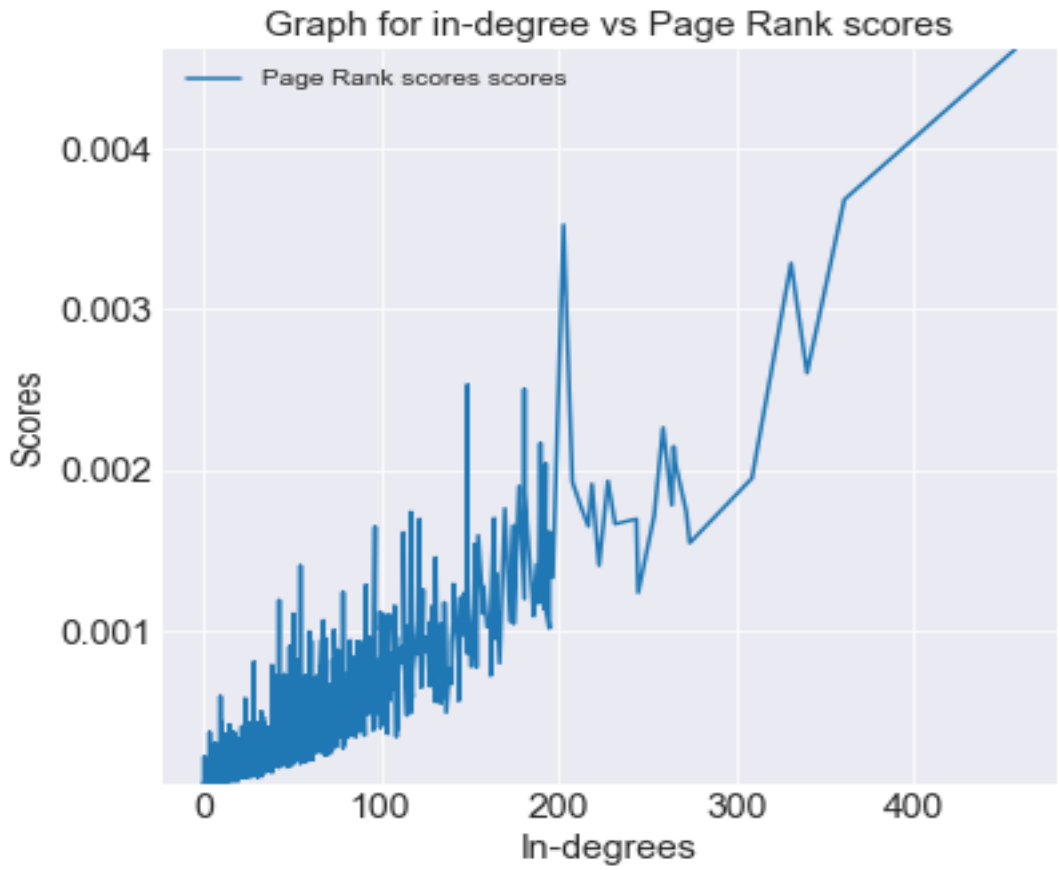
The comparison of the different values is given on the next page:



Inference: As the hub score of a node is obtained by summing up the authority score of all the nodes to which the node has out links. It can be seen from the graph that the node with the highest hub score has the highest outdegree and the hub scores are mostly increasing with the increase in out-degree values.



Inference: As the authority score of a node is obtained by summing up the hub score of all the nodes from which the node has in-links. It can be seen from the graph that the node with the highest authority score has the highest indegree and the authority scores are mostly increasing with the increase in in-degree values.



Inference: It can be seen from the graph that the node with the highest Page rank score has the highest indegree.

Another result which has been found is given below:

- Also, for the top 10 nodes according to hub score if we find the nodes that are the successor of a node as well as present in the list of top 10 authority nodes, we get the below result:

The top 10 hubs along with top 10 authorities that are also their successors

1. 2565

['2398', '4037', '3352', '1549', '762', '3089', '1297', '2625']

2. 766

['2398', '3352', '762', '3089', '2565', '15', '2625']

3. 2688

['2398', '4037', '3352', '1549', '762', '3089', '1297', '2565', '2625']

4. 457

['2398', '4037', '3352', '1549', '762', '3089', '1297', '2565', '15', '2625']

5. 1166

['2398', '4037', '3352', '1549', '762', '3089', '1297', '2565', '15', '2625'

6. 1549

['2398', '4037', '3352', '762', '3089', '1297', '2565', '15', '2625']

7. 11

['3352', '1549', '762', '1297', '2565', '15', '2625']

8. 1151

['2398', '4037', '3352', '1549', '762', '3089', '1297', '15', '2625']

9. 1374

['2398', '4037', '3352', '1549', '762', '3089', '1297', '2565', '2625']

10. 1133

['2398', '4037', '3352', '1549', '762', '3089', '1297', '15', '2625']

So, it can be seen that all the top 10 hubs have some nodes out of top 10 authorities as their successor and that is also one of the reasons these nodes have high hub scores.

- Also, for the top 10 nodes according to authority score if we find the nodes that are the predecessor of a node as well as present in the list of top 10 hub nodes, we get the below result:

  The top 10 Authorities along with top 10 Hubs that are also their predecessors

  1. 2398
  ['2565', '766', '2688', '457', '1166', '1549', '1151', '1374', '1133']

  2. 4037
  ['2565', '2688', '457', '1166', '1549', '1151', '1374', '1133']

  3. 3352
  ['2565', '766', '2688', '457', '1166', '1549', '11', '1151', '1374', '1133']

  4. 1549
  ['2565', '2688', '457', '1166', '11', '1151', '1374', '1133']

  5. 762
  ['2565', '766', '2688', '457', '1166', '1549', '11', '1151', '1374', '1133']

  6. 3089
  ['2565', '766', '2688', '457', '1166', '1549', '1151', '1374', '1133']

  7. 1297
  ['2565', '2688', '457', '1166', '1549', '11', '1151', '1374', '1133']

  8. 2565

['766', '2688', '457', '1166', '1549', '11', '1374']

9. 15
['766', '457', '1166', '1549', '11', '1151', '1133']

10. 2625
['2565', '766', '2688', '457', '1166', '1549', '11', '1151', '1374', '1133']

So, it can be seen that all the top 10 authorities have some nodes out of top 10 hubs as their predecessor and that is also one of the reasons these nodes have high authority scores.

- Also, for the top 10 nodes according to PageRank score if we find the nodes that are the predecessor of a node as well as present in the list of top 10 PageRank nodes, we get the below result:

  The top 10 Page Rank nodes along with top 10 Page rank nodes that are also their predecessors

  1. 4037
  ['15']

  2. 15
  ['4037', '2237']

  3. 6634
  []

  4. 2625
  ['2398']

  5. 2398
  ['15', '2237', '4191', '5254']

  6. 2470
  ['2237']

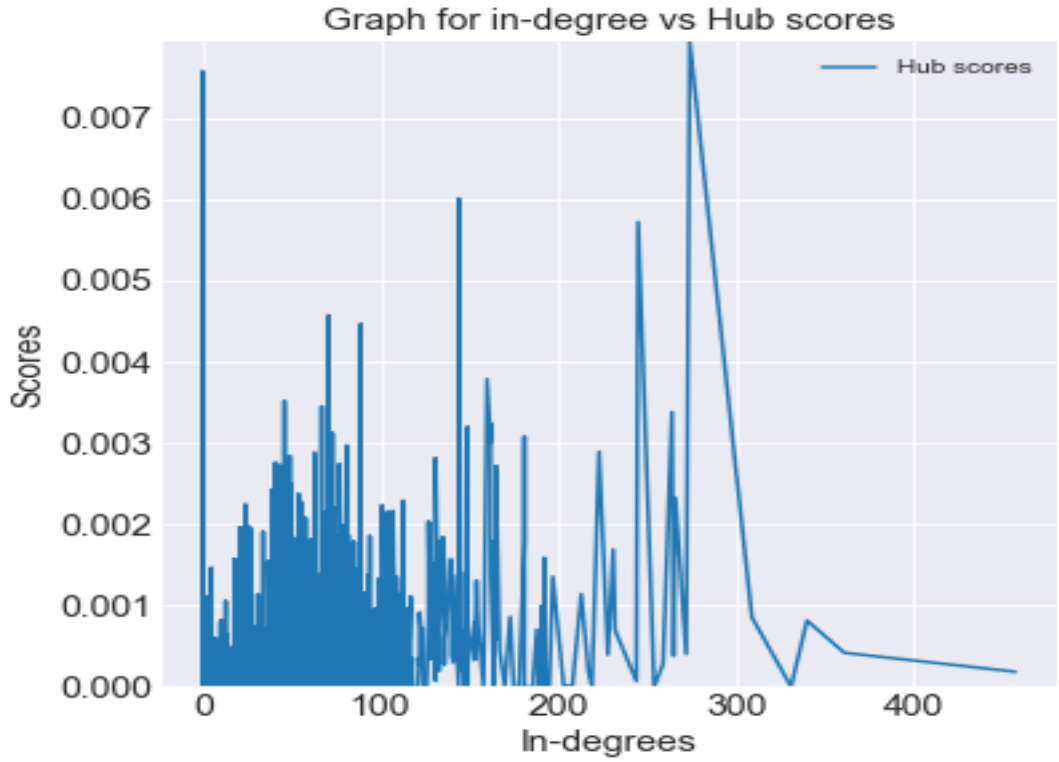
  7. 2237
  []

  8. 4191
  ['2398', '2237', '5254']

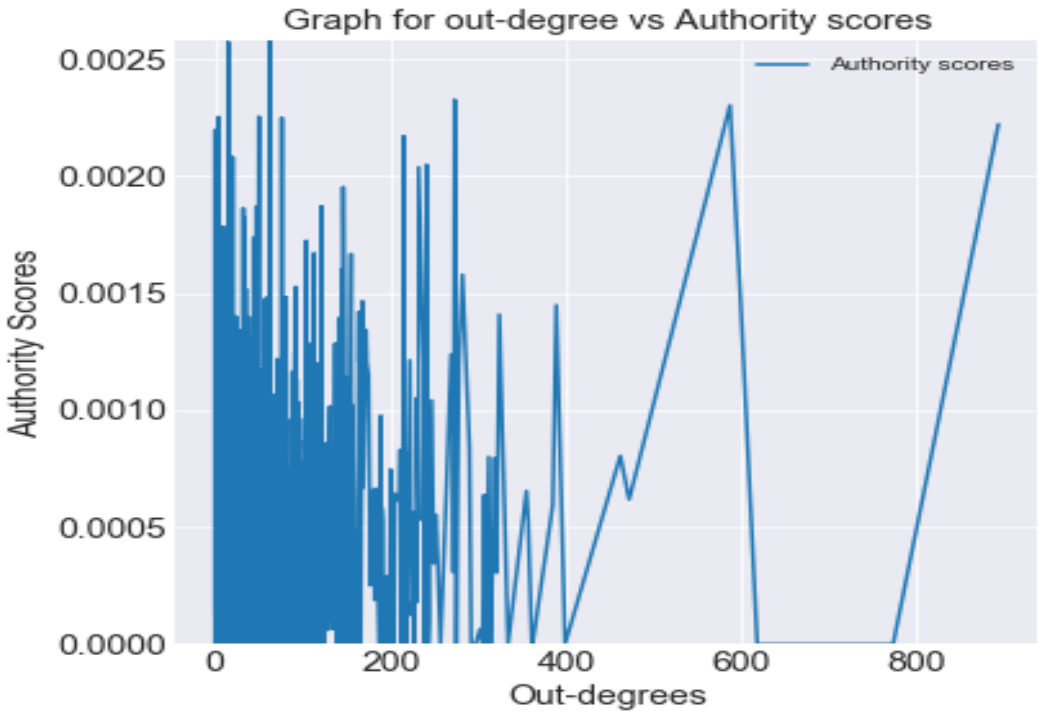
  9. 7553
  []

  10. 5254
  []

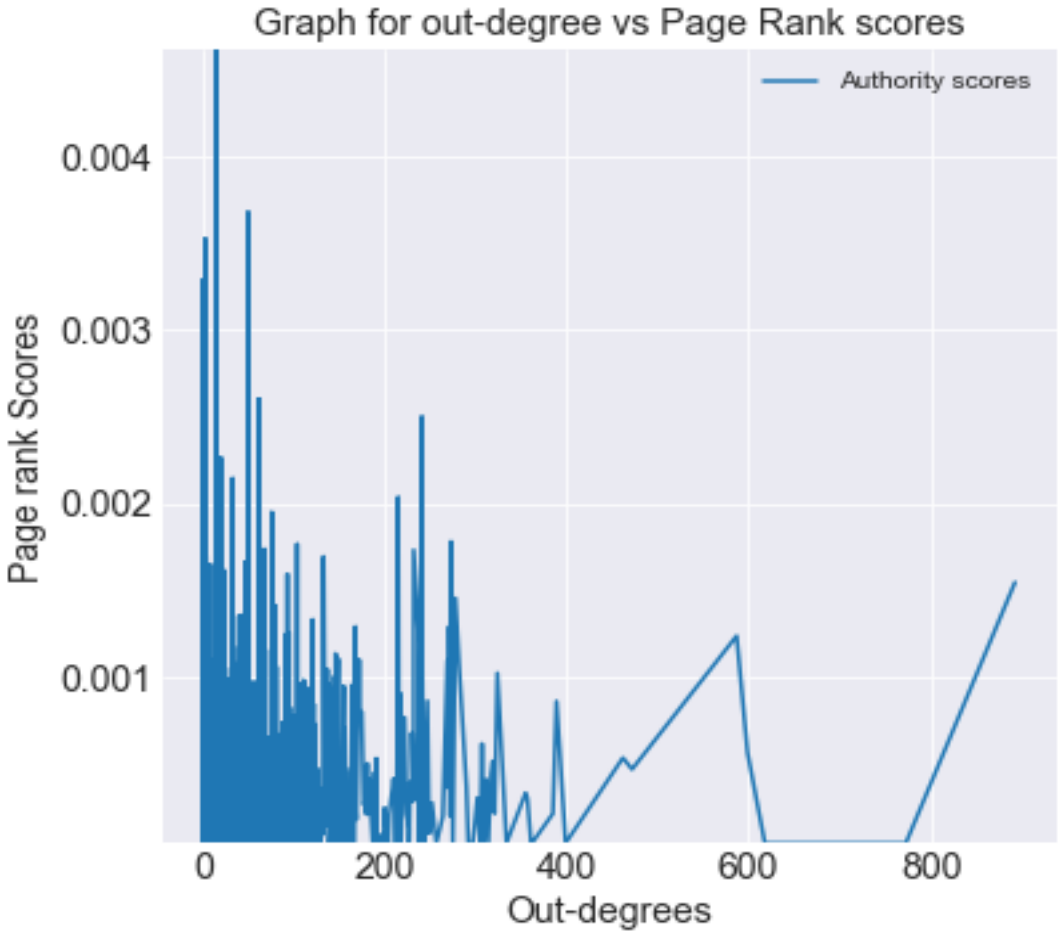
So, it can be seen most of the top 10 PageRank nodes have some nodes out of top 10 PageRank nodes as their predecessor and that is also one of the reasons these nodes have high PageRank scores.



Inference: Hub score has no relation with indegree of the node.



Inference: Authority score has no relation with out-degree of the node.



Inference: PageRank score mostly has no relation with out-degree of the node.