Analysis Assi 6

-Shalini Bhardwaj(MT19045)

Ques 1

Data set Used: CollegeMsg

A directed graph is used in both the questions.

Intially adjajency list of edges and nodes in build and then 2D adjacency matrix is built to calculate below information about graph .

1.Total number of Nodes and edges came out to be:

```
Total nodes ........ 1899
Total edges ........ 20296
```

2. Average indegree =sum of indegree of all nodes/Total nodes

Average outdegree =sum of indegree of all nodes/Total nodes

Maximum Indegree and Outdegree are displayed below

Density of Network= Total No.of edges/(Total Nodes)*(Total Nodes -1)

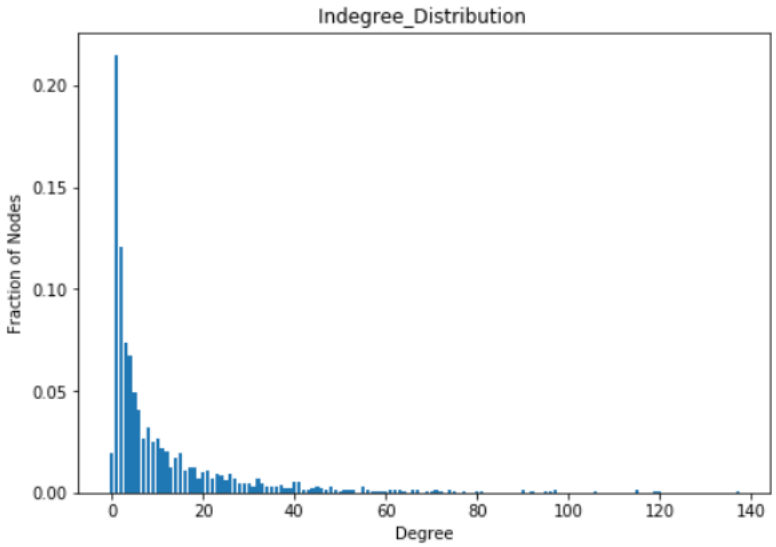
```
AVG_indegree......... 10.687730384412848
Max Indeg node.... 32 ...... with degree 137
AVG_outdegree......... 10.687730384412848
Max Outdeg node.... 9 ...... with degree 237
Density_network.... 0.005631048674611617
```

2.Degree Distribution

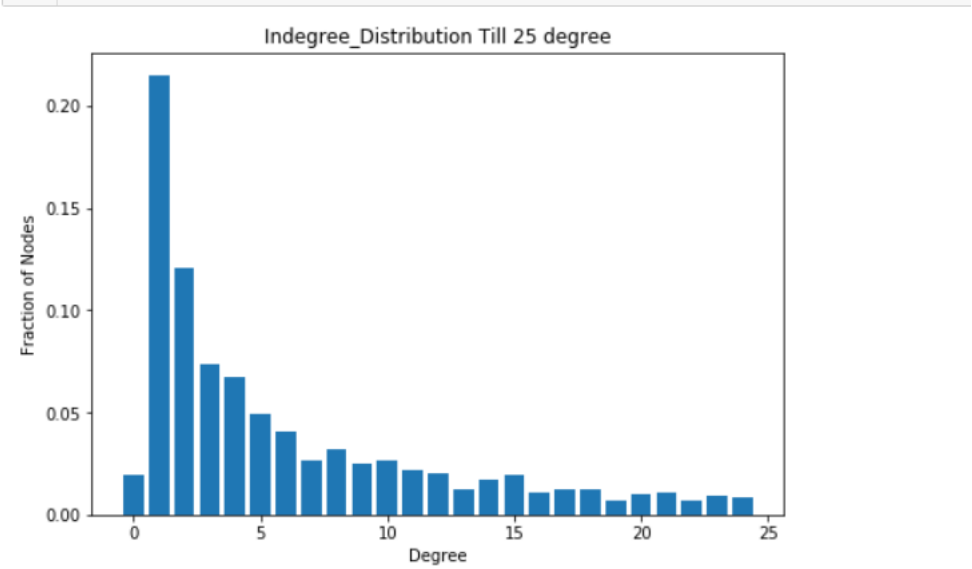
Indegree distribution

On y axis Fraction of nodes =

number of nodes with in degree/Total Number of Nodes



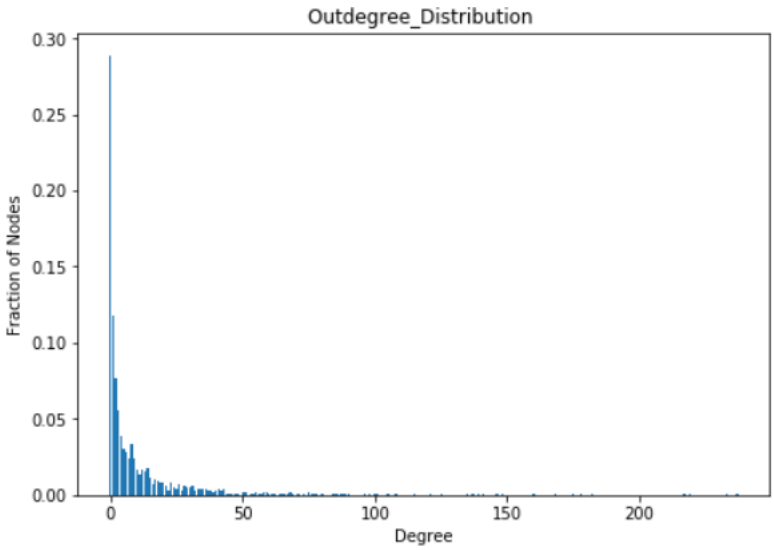
In degree Distribution for 50 nodes for clarity



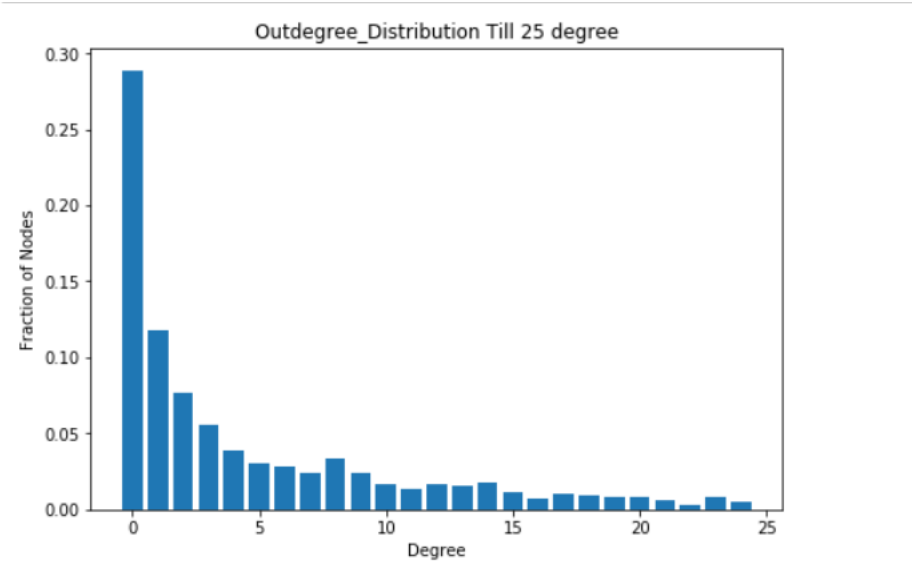
Outdegree distribution

On y axis Fraction of nodes =

number of nodes with out degree/Total Number of Nodes



Out degree Distribution for 50 nodes for clarity

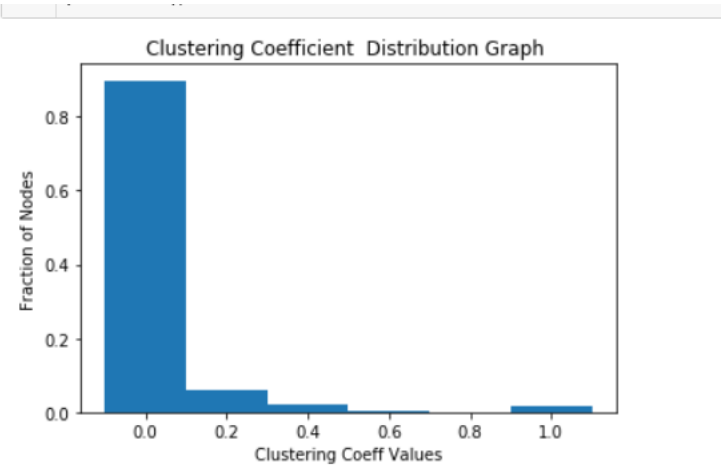


3. Clustering coefficient

Clustering Coefficient for each node is calculated as :

Total edges between neighbours of node/ Number of neighbors* (Number of neighbors -1)

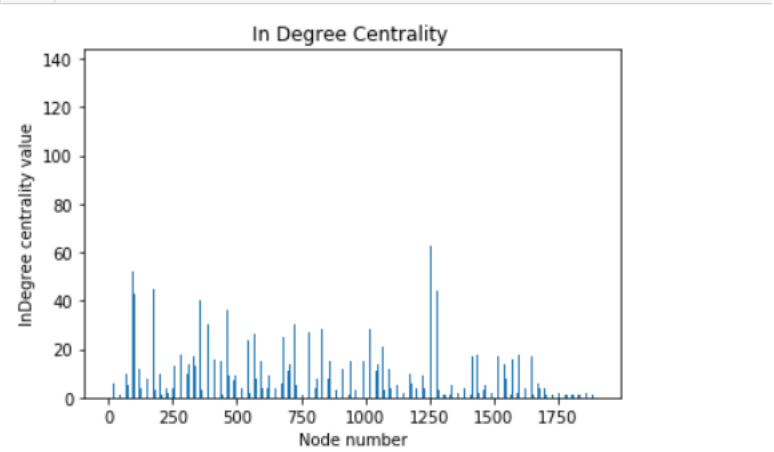
It varies between 0 and 1.



4.Centrality measure Used: Degree centrality

Here both indegree and out degree centrality distribution is built.

Degree Centrality shows how much indegree/outdegree is there for each node.For directed graph it is usually seen indegree is important.



5. Page Rank Score calculated

Page rank algorithm uses **only indegree of node** as metric to rank pages.

Google uses this algorithm.It is belived that a page if is highly prestigious and valuable will eventually haver high indegree.Therfore it do not consider out degree of nodes.

Advantages:

1. PageRank can be precomputed, HITS has to be computed at query time.
2. HITS is too expensive in most application scenarios.
3. Less susceptible to localized link-spam
4. As per page rank node 32 has highest pr score node. It is the node of highest indegree.
   Result of Top 50 pr score node is shown:

```
page rank score
32 : 0.005997514282124592
42 : 0.005897543015508087
638 : 0.005389242638293347
372 : 0.005086009601527966
400 : 0.0045385134983739394
103 : 0.004415041936696585
598 : 0.004388769134484078
194 : 0.004194510126481766
713 : 0.0038701731196683544
249 : 0.003869590427168905
325 : 0.0038351551282263976
1283 : 0.003825216947003185
840 : 0.003642391924095981
72 : 0.0034680511561095205
97 : 0.0034028469382127707
105 : 0.0033013505545111733
212 : 0.0031683652124096006
1624 : 0.003113690635683903
308 : 0.003061476041859718
67 : 0.0030345698527599074
1189 : 0.0030109193277403344
475 : 0.0029911742403621773
128 : 0.002900581151254372
63 : 0.002896124847548514
644 : 0.002838224222718721
1713 : 0.002825188744788339
191 : 0.002707718524036158
697 : 0.002680603239501101
542 : 0.0026785520378683385
254 : 0.0026581018412772254
513 : 0.002643538165114415
263 : 0.0026188755796014104
```

Thus in some cases it is seen pr score is directly propotional to indegree.

6.Authority Score

1. It is seen high prestige implies good authority.Authorirty score is calculated in Hits.It measures the indegree just like page rank.
2. Authority score of a node is sum of hub score of its neighbour.
3. Highest indegree node can be seen in top authority score node.
4. Top 50 nodes with high authority score is printed:

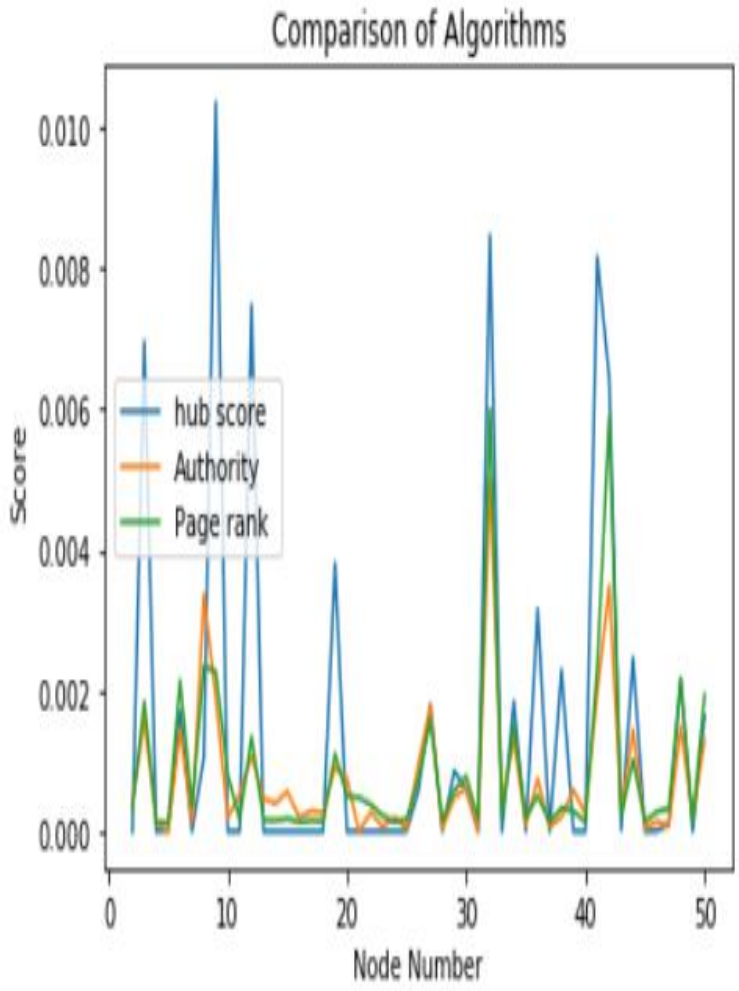
```
authority score
598 : 0.005920012139941655
32 : 0.005004856937386899
638 : 0.0048580786965861005
840 : 0.00421373771151018
713 : 0.004206635713856448
475 : 0.003962435311005718
72 : 0.003786698889869976
97 : 0.0036216057275247453
502 : 0.0035696287039057146
626 : 0.003568271337299199
194 : 0.0035435297732044073
1255 : 0.003538974761336863
372 : 0.00352455390758928 1
42 : 0.0034906045433238764
8 : 0.003367084698386315
308 : 0.003335065043449349
297 : 0.003268644772303222
63 : 0.0032552811799298627
```

7. Hub Score

1. Hub score is measured on the out degree of node.
2. It is computed as sum of authority score of its neighbour.
3. It is used in Hits algorithm which uses both in degree and out degree.
4. Here node 9 has maximum outdegree and is coming in top position in ranking.
5. This shows hub incorporates outdegree.

```
hubscore
103 : 0.011197129953017648
9 : 0.010362488921315133
105 : 0.01034654749678217
400 : 0.00913147237997526
249 : 0.008922464333577015
32 : 0.0084740212491942
41 : 0.008165782116625192
357 : 0.007630927780623781
12 : 0.00748189538869325
67 : 0.007051755591164902
3 : 0.0069598274859080074
372 : 0.006902052260427435
42 : 0.006450523282931995
1283 : 0.006149256498730235
770 : 0.005988738347690775
194 : 0.005849940064512231
176 : 0.005752265169876248
321 : 0.005640454207391711
1713 : 0.005388297716075327
638 : 0.005273038713260522
1189 : 0.00525127748192153
713 : 0.004937566899813196
704 : 0.004685855881052049
431 : 0.004483446923681091
325 : 0.004340958186053091
733 : 0.004289517824380027
1281 : 0.004150940177670415
```

Ranking Algorithm Comparison for 50 nodes



1.Here we can see as both page rank and authority rank on the basis of indegree and thus have very similar behaviour .Orange and Green lines in above Graph.

2.Hub score is showing varied behaviour compared to bith authority score and page rank.It can be seen in blue color.

3.Above Graph is built on 50 nodes to clearly show the variation.

4.Blue line is showing spike at node 9 as it is node with maximum outdegree in dataset.

5.Orange and green line is showing spike at node 32 as it is node with maximum indegree in dataset.

Page Rank Advantages:

1.Low computation cost as PageRank can be precomputed, HITS has to be computed at query time.

2. Google uses it as Google uses this algorithm.It is belived that a page if is highly prestigious and valuable will eventually haver high indegree.Therfore it do not consider out degree of nodes.

HITS Advantages:

1.Takes both indegree and outdegree in consideration.

2.More topic/query sensitive results.