REPORT

Tools Used :

- Pandas
- Matplotlib
- Numpy
- Networkx

Data Set Used :

Name : email-Eu-core network

Link :https://snap.stanford.edu/data/email-Eu-core.html

Description : The network was generated using email data from a large European research institution. There is an edge (X-Y) in the graph if a person X has sent an email to another person Y within the institution.The graph is a directed graph.The dataset also contains "ground-truth" community memberships of the nodes. Each individual belongs to exactly one of 42 departments at the research institute.

Question 1 :

Creating Adjacency Matrix and Edge List

I have created a column stochastic adjacency matrix.The shape of the matrix is 1005 X 1005.

```
[[1 0 0 ... 0 0 0]
 [1 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

The edge is constructed from the adjacency list using defaultdict defined in python's collection library.

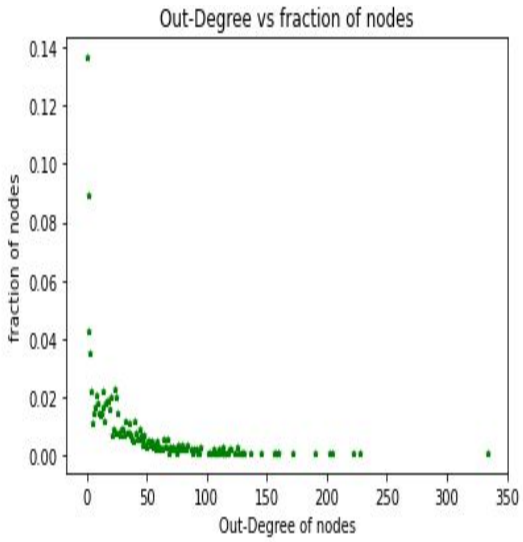
```
defaultdict(<class 'list'>, {0: [0, 1, 5, 6, 17, 18, 64, 73, 74, 88, 101,
103, 146, 148, 166, 177, 178, 215, 218, 221, 222, 223, 226, 238, 248, 250,
266, 268, 283, 297, 309, 313, 316, 368, 377, 380, 459, 498, 560, 581,
```

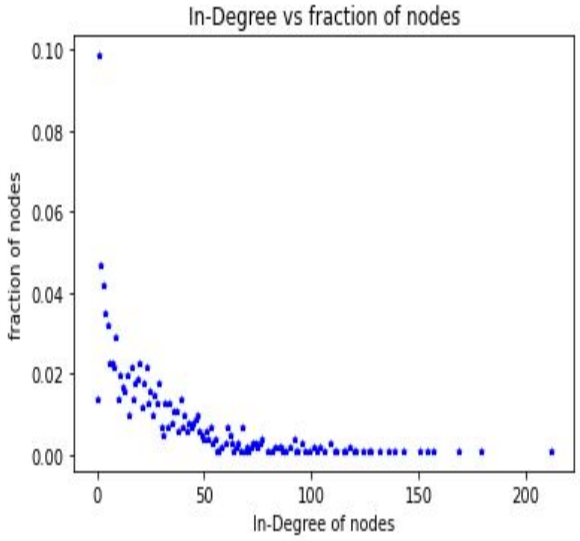
734], 1: [1], 2: [2, 3, 4, 5, 6, 13, 54, 55, 56, 57, 58, 59, 63, 64, 86, 89, 9…….

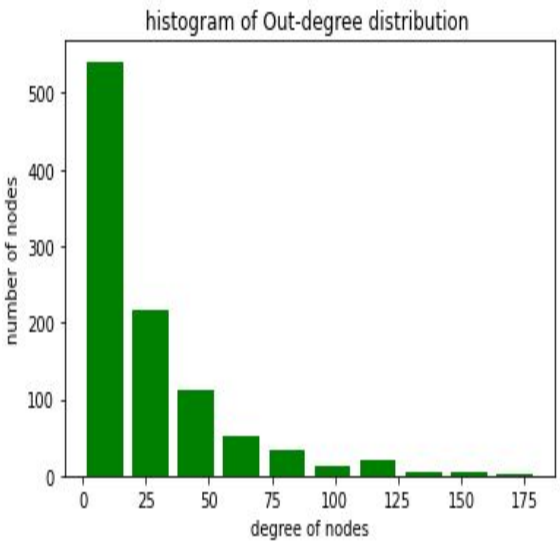
Dataset Statistics	Value
Number of Nodes	1005
Number of Edges	25571
Avg In-degree	25.4437
Avg Out-degree	25.4437
Node with Max In-degree	160
Node with Max Out-degree	160
Density of the network	0.0253171

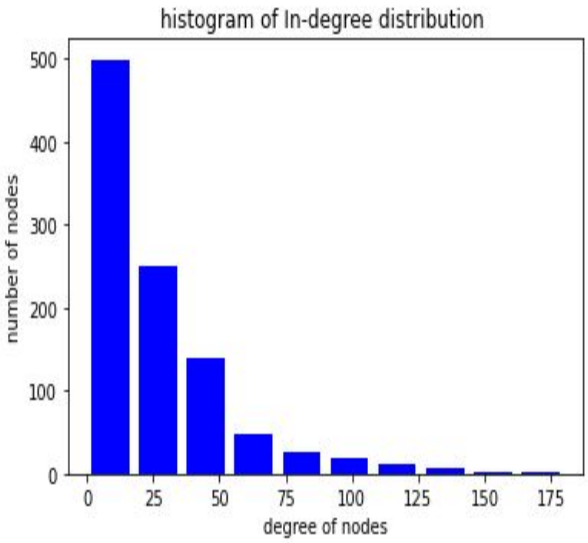
Degree Distribution Of The Network

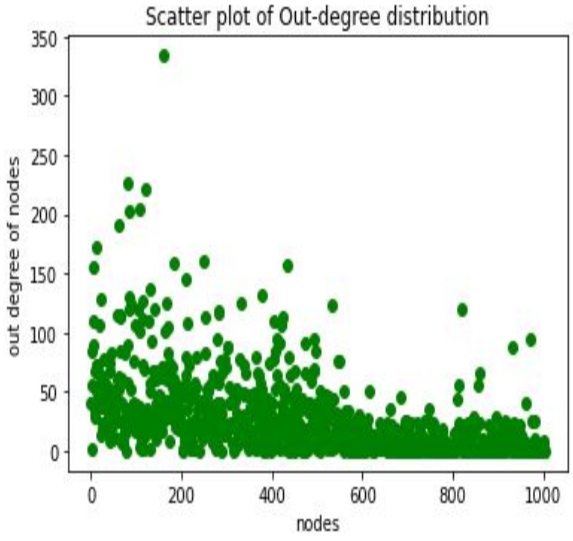
I have plotted below curves to find the degree distribution of the network and draw inferences.

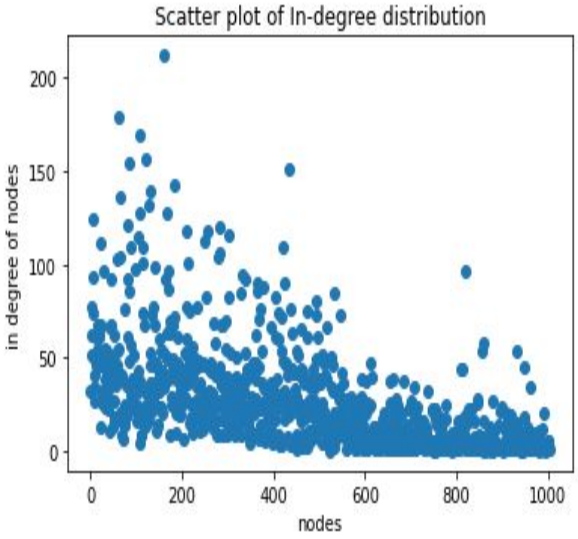












Inferences Drawn

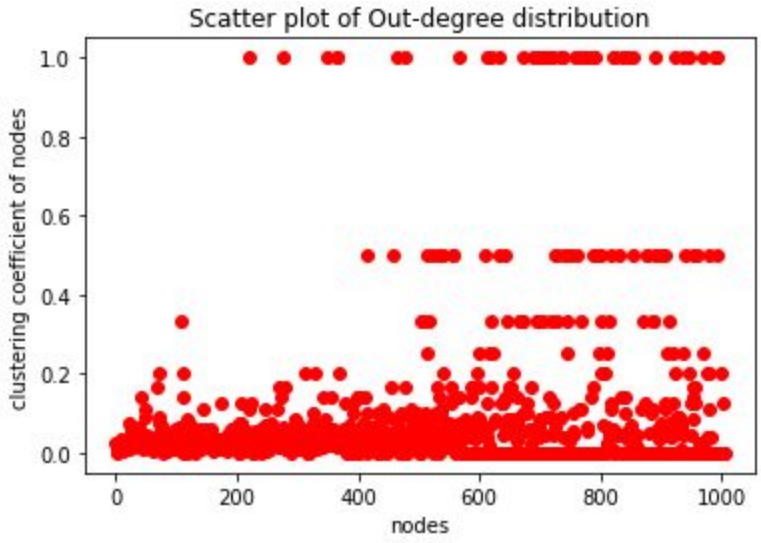
- Network is sparse as there is a high fraction of nodes that have indegree and out degree equal to zero.

- Node 160 seems to be an important node as the graph says that it has the highest in degree and out degree.

- There is a smooth declination in the number of nodes with increasing in degree.The same pattern is observed for the out degree as well but the declination is not very smooth.

- The highest indegree of any node is less than the highest out degree of any node in the network which suggests that all the nodes are not of equal importance.

Clustering Coefficient Of Nodes



Following formula is used to this directed network :

V : Node
Kv : out - degree of the node
Nv : Number of links between neighbors of V

Clustering Coefficient = Nv / Kv *(Kv-1)

Inferences Drawn :

- Average clustering coefficient of the network is 0.10562845.

- The clustering coefficient equal to one denotes that there already exists triadic closure between nodes. There are 43 nodes in the network with clustering coefficient equal to one.These all nodes are forming a complete dense cluster.

- Similarly there are 35 nodes with clustering coefficient = 0.5 .These nodes also tend to create tightly knit groups characterised by a relatively high density of ties.

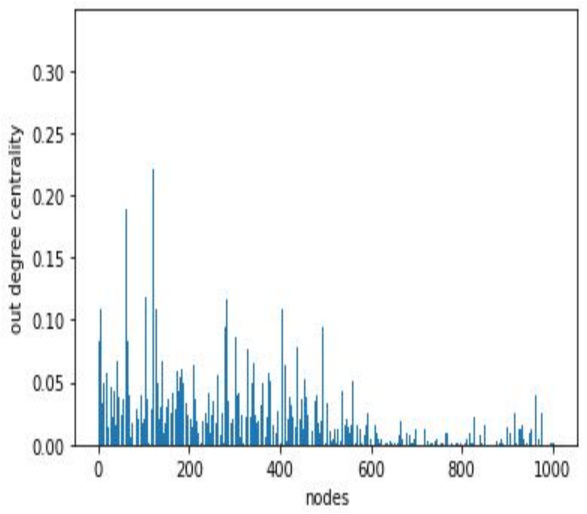
- There are a good amount (227) of nodes with zero clustering coefficient.Which means they are individual nodes not part of any relationship with other nodes in the network.

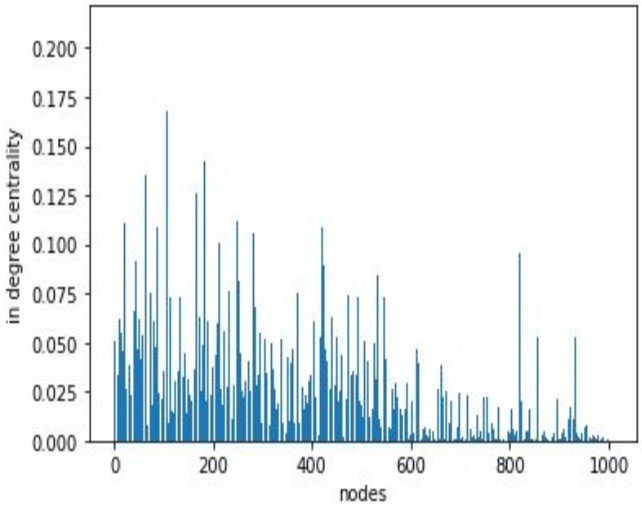
Centrality Measure Chosen : Degree Centrality

Degree Centrality metric is the importance of a node in the network as being measured based on its degree i.e the higher the degree of a node, the more important it is in a graph

Ref :
https://networkx.github.io/documentation/networkx-1.9/reference/generated/networkx.algorithms.centrality.degree_centrality.html

The degree centrality for a node v is the fraction of nodes it is connected to.Have reported in degree centrality as well as out degree centrality.





Statistics Name	Value
Top 10 Nodes with highest outdegree centrality	[434, 183, 249, 13, 62, 86, 107, 121, 82, 160]
Top 10 Node with highest indegree centrality	[128, 64, 129, 183, 434, 86, 121, 107, 62, 160]
Average outdegree centrality	0.025317
Average indegree centrality	0.025317

Inferences Drawn :

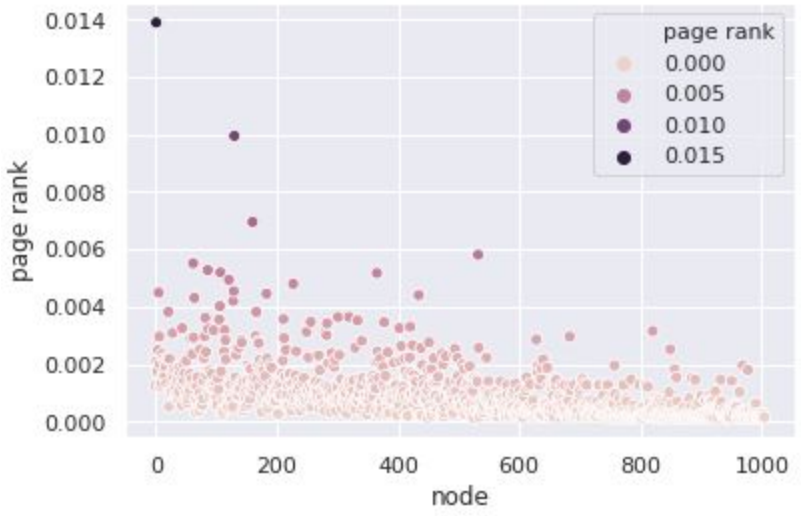
- High In degree centrality means that node is of high importance.This means that top 10 nodes with highest in degree centrality mentioned above are high rank officers of the institution.These are the most influential people of the institution.

- Node number 160 seems to be the director of the research institute as this node has highest indegree and out degree centrality.

- Top 10 nodes with highest out degree centrality may belong to the research students of the institution who are following their instructors and other important folks.

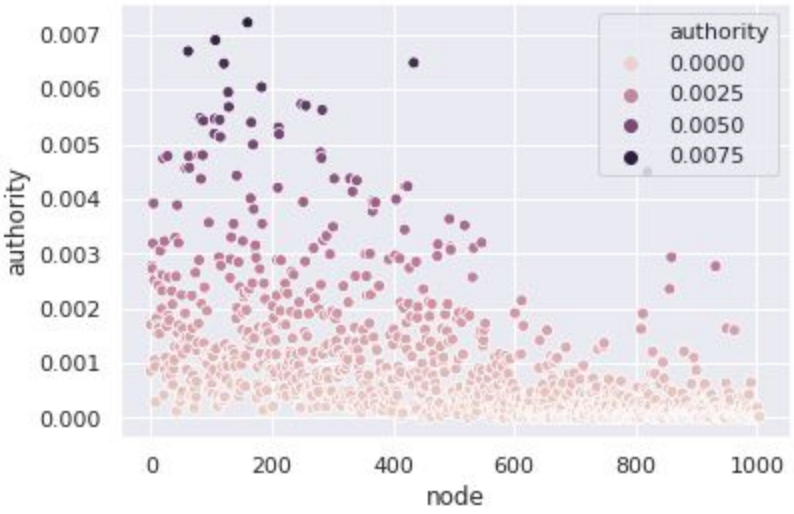
Question 2

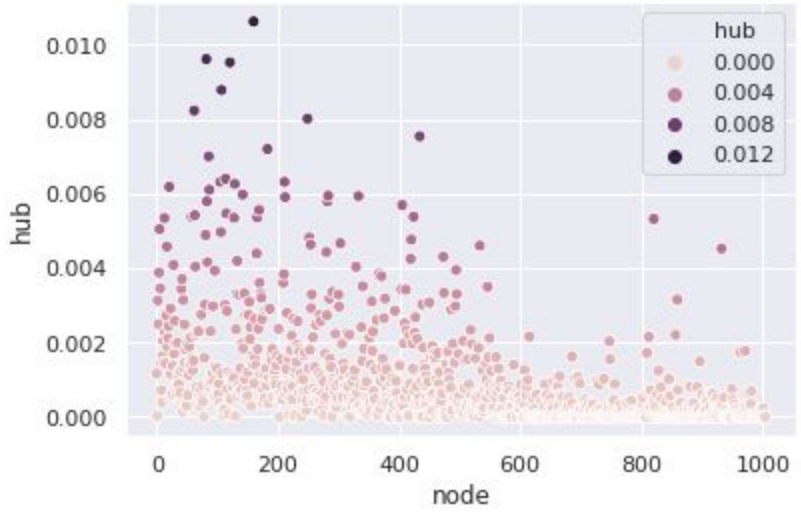
Page Rank Distribution of the network



Statistics Name	***Value***
Nodes with top 10 page ranks	[227, 121, 365, 107, 86, 62, 532, 160, 130, 1]
Highest Page Rank	*0.01389620*
Node with Highest Page Rank	1

Authority & Hub Score of the network nodes





Statistics Name	Value
Highest Authority Score	0.0072204
Highest Hub Score	0.0106288
Nodes with top 10 authority score	[129, 256, 249, 128, 183, 121, 434, 62, 107, 160]

Nodes with top 10 hub score	[114, 86, 183, 434, 249, 62, 107, 121, 82, 160]

Inferences :

- Node having the highest hub score and authority score are the same i.e. node number 160.
- Good hub almost always is good authority as per the scatter plots drawn above.
- Nodes belonging of top 10 page rank , authority score and hub score are overlapping.

Comparison in the results obtained from Page Rank vs HITS

Relevance :

- HITS relevance is less since the algorithm ranks the pages on the indexing time whereas Page Rank results are more relevant as the algorithm uses the hyperlinks to give good results and takes page content into consideration.
- Top 10 nodes by page Rank : [227, 121, 365, 107, 86, 62, 532, 160, 130, 1]
- Top 10 pages by Authority : [129, 256, 249, 128, 183, 121, 434, 62, 107, 160]
- page number 160 , 121 , 107 , 365 are top 10 pages ranked by both the alogs. Therefore are the most relevant pages of the network.
- Remaining pages drawn by Authority score i.e 129 , 256 , 249 , 183, 434 are less relevant as compared to the remaining pages drawn by Page Rank.

Neighborhood :

HITS is applied to the local neighborhood of nodes surrounding the results of a query whereas Page Rank is applied to the whole WWW structure. Therefore the nodes which are having high Authority score are localised to a particular area whereas the nodes resulted by Page Rank are global results.

Overlapping Page Rank and HITS Results :

- All the nodes which are having high Page rank and authority score are the one on the high posts of this research institute.

- In our network the node 160 is the node with top 10 hub and authority score.The page rank for this node is 3 in the network.This highly suggests that this is one of the most important people of the institute.

- Nodes having good hub scores are having good authority scores as well. These nodes are also ending up with high page ranks as per observed in the scatter plots drawn above.

- Page Rank suffers from the problem of Rank Sinks, Spider Traps and etc Dangling Links.Similarly HITS suffers from Mutually reinforcing relationships between hosts and Topic Drift.This could lead to give less importance to few nodes which are having higher rank in the institution.