Report(Assignment 3)

Question 1:

The dataset that we used is: **email-Eu-core network**

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

This dataset contains directed graph in edge list format

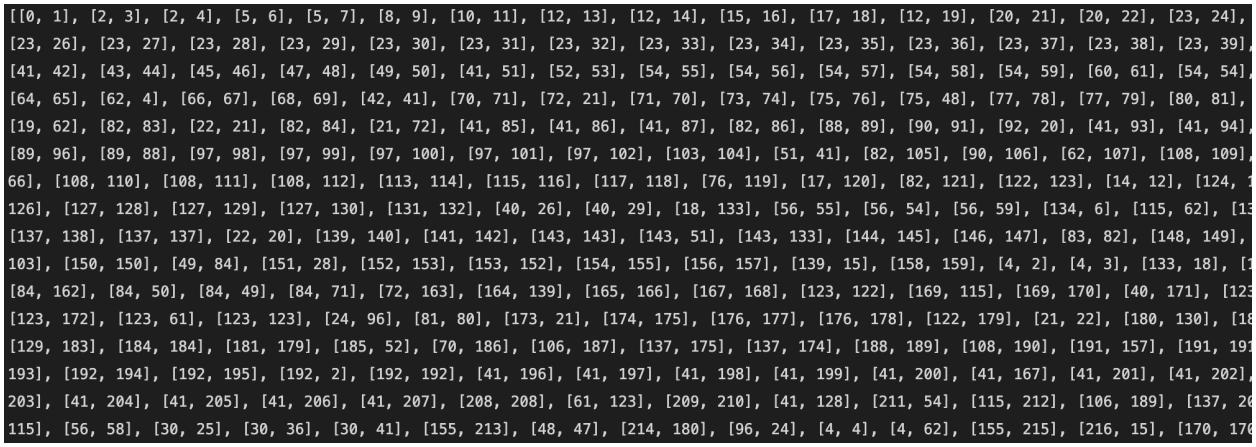
This dataset contains the network that was generated using the email data from an institution based in Europe. The edge(u,v) describes the fact that the person u sent person v atleast one email.

Dataset statistics	
Nodes	1005
Edges	25571
Nodes in largest WCC	986 (0.981)
Edges in largest WCC	25552 (0.999)
Nodes in largest SCC	803 (0.799)
Edges in largest SCC	24729 (0.967)
Average clustering coefficient	0.3994
Number of triangles	105461
Fraction of closed triangles	0.1085
Diameter (longest shortest path)	7
90-percentile effective diameter	2.9

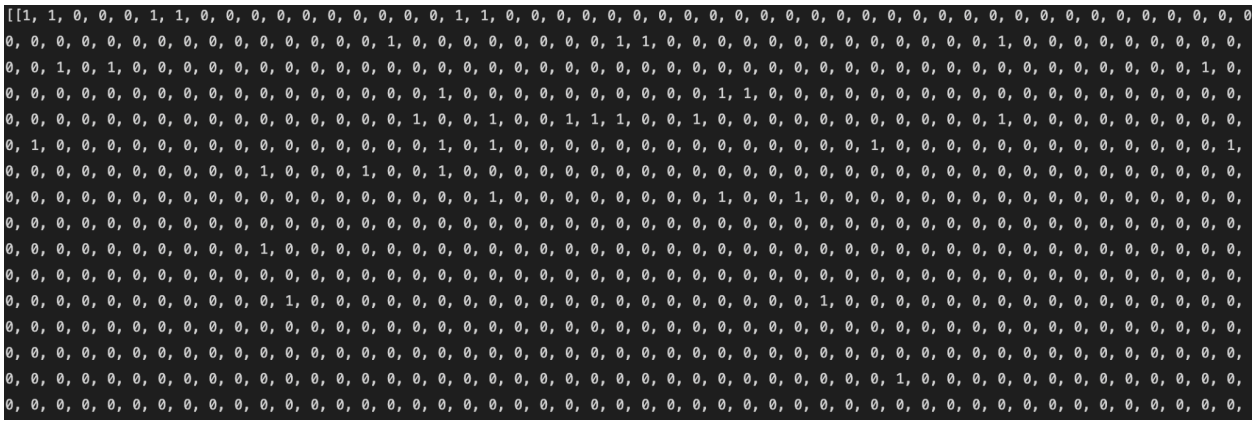
For the first part, first we extracted the data, and then formed the edge list by appending the edges in a list and also made a list containing all the nodes.

```
Number of nodes is  1005
Number of edges is  25571
average in-degree is  25.443781094527363
max in-degree is  212
average out-degree is  25.443781094527363
max out-degree is  334
Network density is  0.025342411448732432
```

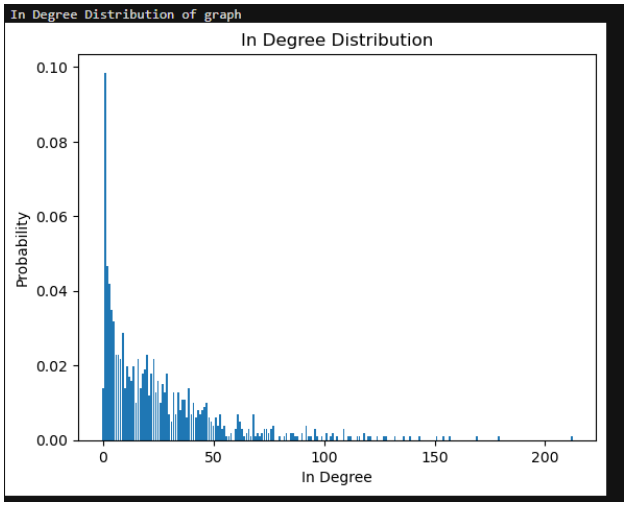
The edge list



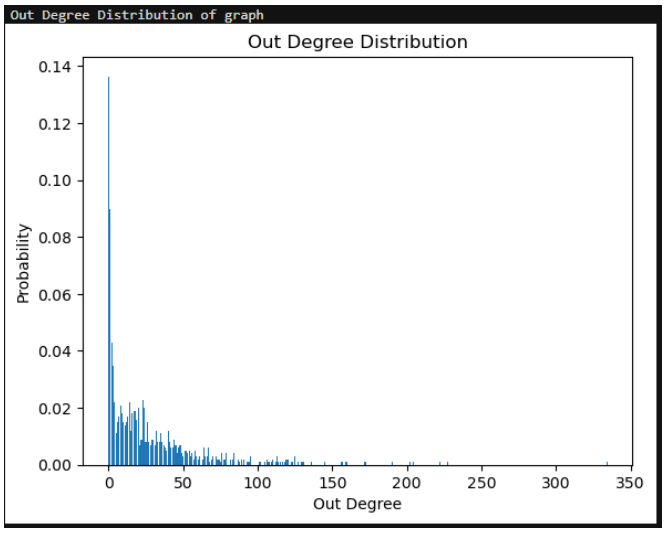
The adjacency matrix:



(i)The following barplot denotes the indegree(the number of inward directed graph edges) distribution of nodes in the chose graph:



(ii)The following barplot denotes the outdegree(the number of outward directed graph edges from a given graph vertex in a directed graph.) distribution of nodes in the chose graph:



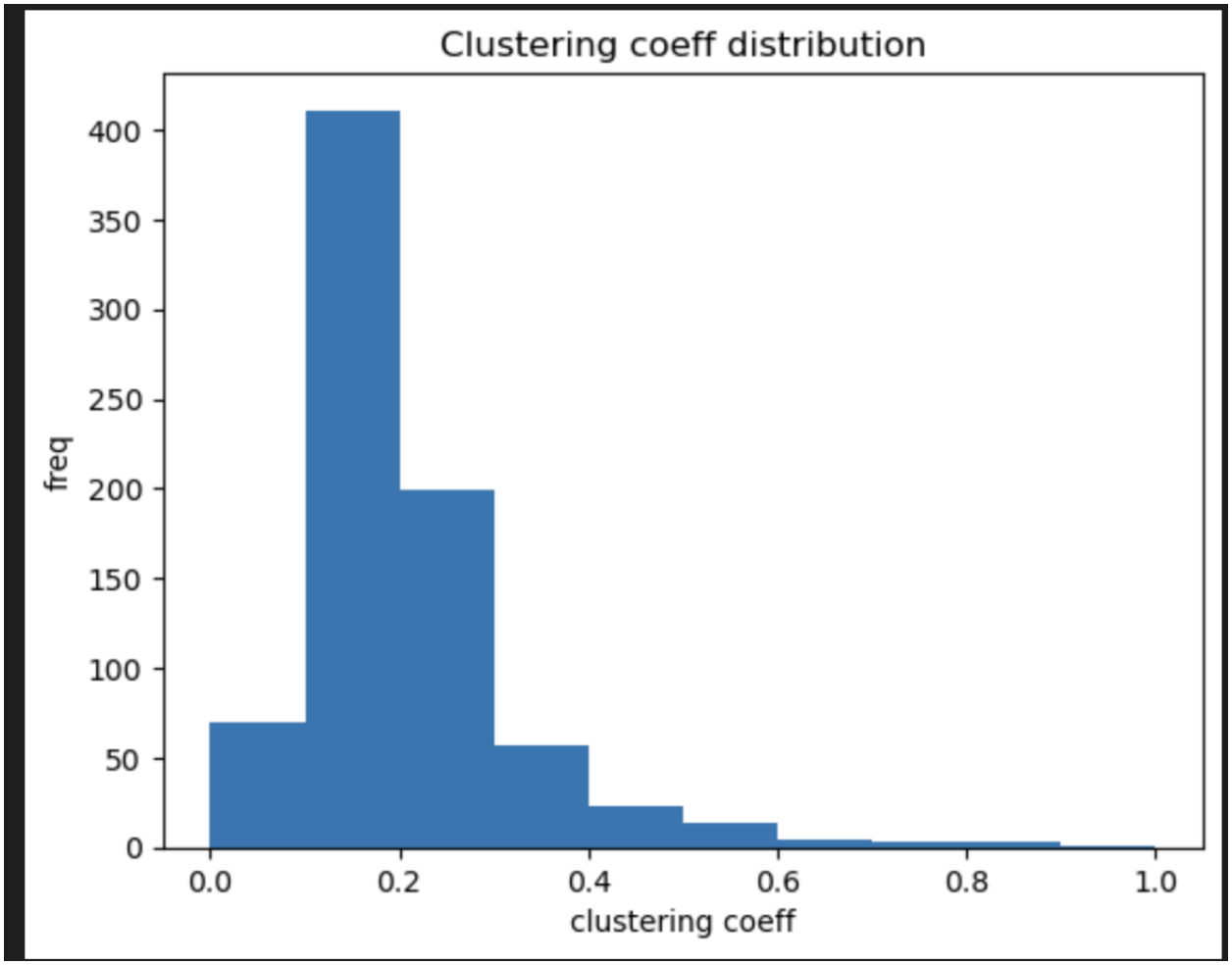
These two graphs collectively give us a picture of the sparsity of the graph and the link density between the nodes. As can be seen, the highest number of nodes have a very less indegree and outdegree, thus indicating a sparse distribution.

Clustering Coefficient :
Formula used -> clustering coeff of a node = 2*T/n*(n-1) where T is the number of Triangles which the node is part of and n is the number of neighboring nodes of current node

Clustering coefficients:

```
clustering coefficient for node 0 is 0.11288685946220194
clustering coefficient for node 1 is 0.16339869281045752
clustering coefficient for node 2 is 0.11420807453416149
clustering coefficient for node 3 is 0.1521802115022454
clustering coefficient for node 4 is 0.11956878487212502
clustering coefficient for node 5 is 0.04532770097286226
clustering coefficient for node 6 is 0.06149122374924059
clustering coefficient for node 7 is 0.11381809095452274
clustering coefficient for node 8 is 0.17380574651859687
clustering coefficient for node 9 is 0.17250470809792842
clustering coefficient for node 10 is 0.11537537537537539
clustering coefficient for node 11 is 0.08210825571391164
clustering coefficient for node 12 is 0.12021198830409356
clustering coefficient for node 13 is 0.05710744413201125
clustering coefficient for node 14 is 0.08979248687664042
clustering coefficient for node 15 is 0.14679313459801266
clustering coefficient for node 16 is 0.115255530129672
clustering coefficient for node 17 is 0.11002092201139889
clustering coefficient for node 18 is 0.12142981539533264
clustering coefficient for node 19 is 0.10534547152194211
clustering coefficient for node 20 is 0.11091795837558549
clustering coefficient for node 21 is 0.08300209205020921
clustering coefficient for node 22 is 0.17426400759734093
clustering coefficient for node 23 is 0.13017580383992597
clustering coefficient for node 24 is 0.1626557433009046
...
clustering coefficient for node 1003 is 0
clustering coefficient for node 1004 is 0
```

Clus_coeff_distribution:



It can be seen that most nodes have clustering coeff around 0.2

Question 2:

In Question 2, to calculate the pagerank,hub, authority score networkx API of Python was used. First, we added all the nodes and edges in the directed graph we initialized. Then we use the pagerank function of networkx to calculate the pagerank of all the nodes.

```
pagerankscores = nx.pagerank(Graph)    #calculates the p


for n, pg_score in pagerankscores.items():
    print("Node" ,n, ": PageRank score -",pg_score)
```

Similarly, to calculate the hub score and authority score, networkx.hits was used. The hub score is calculated based on the outgoing links and the authority score is calculated based on the incoming links.

```
hubscores = nx.hits(Graph)[0]                    #calculates the hub score for al
authorityscores = nx.hits(Graph)[1]              #calculates the authority score


for n, hub_score in hubscores.items():
    a_score = authorityscores[n]
    print("Node" ,n, ": Hub score -", hub_score, ", Authority score -", a_score)
```

COMPARISON BETWEEN THE RESULTS OBTAINED IN PART 1 AND PART 2

```
Node 0 : PageRank score - 0.0012754775504594832
Node 1 : PageRank score - 0.009411560186382712
Node 2 : PageRank score - 0.002095369381767161
Node 3 : PageRank score - 0.0017234388700051611
Node 4 : PageRank score - 0.002436232111121507
Node 5 : PageRank score - 0.004525470848399022
Node 6 : PageRank score - 0.0029192068825005272
Node 7 : PageRank score - 0.0019464642315471185
Node 8 : PageRank score - 0.0012364305542142788
Node 9 : PageRank score - 0.001212215032480912
Node 10 : PageRank score - 0.0013292723987216259
Node 11 : PageRank score - 0.002289046520387193
Node 12 : PageRank score - 0.0016365498076585522
Node 13 : PageRank score - 0.002174677080911033
Node 14 : PageRank score - 0.0017008199595549105
Node 15 : PageRank score - 0.001589802529811306
Node 16 : PageRank score - 0.0020210102759722526
Node 17 : PageRank score - 0.001907055205270155
Node 18 : PageRank score - 0.0013262819820475404
Node 19 : PageRank score - 0.0021196946823522694
Node 20 : PageRank score - 0.0022199973212970563
```

output for part 1 (pagerank)

```
Node 0 : Hub score - 0.0011656523581661017 , Authority score - 0.0008525503876603124
Node 1 : Hub score - 2.9004622902639827e-05 , Authority score - 0.0017024061489624475
Node 2 : Hub score - 0.003125339506305258 , Authority score - 0.0027787612230178427
Node 3 : Hub score - 0.0024907646193905015 , Authority score - 0.002723850949124369
Node 4 : Hub score - 0.0038790281781767 , Authority score - 0.003186003461593549
Node 5 : Hub score - 0.005048254863291743 , Authority score - 0.003917430684266547
Node 6 : Hub score - 0.0034511420631543363 , Authority score - 0.002512986894272629
Node 7 : Hub score - 0.0014670269392468944 , Authority score - 0.0010488209777385652
Node 8 : Hub score - 0.0007134118810922484 , Authority score - 0.0009345511864358774
Node 9 : Hub score - 0.00038778628948477663 , Authority score - 0.0002884959959336264
Node 10 : Hub score - 0.0016319219935539183 , Authority score - 0.0018150659893945188
Node 11 : Hub score - 0.0020545042180840894 , Authority score - 0.001541943249067599
Node 12 : Hub score - 0.001465790857041766 , Authority score - 0.0016786629817342408
Node 13 : Hub score - 0.0053450072742758595 , Authority score - 0.002425476508237027
Node 14 : Hub score - 0.0022753271340566392 , Authority score - 0.0016370612677592815
Node 15 : Hub score - 0.0014345706495043512 , Authority score - 0.0015378836679498086
Node 16 : Hub score - 0.002732362557271792 , Authority score - 0.003051574332558066
Node 17 : Hub score - 0.004576554119562811 , Authority score - 0.0026352411911118655
Node 18 : Hub score - 0.0018919144873631518 , Authority score - 0.001993106948160757
Node 19 : Hub score - 0.0021138657400882814 , Authority score - 0.002322438505944743
Node 20 : Hub score - 0.002436209700633867 , Authority score - 0.002606831536911255
```

Output for part 2 (hub and authority)

EXPLANATION:

Pagerank score of a node indicates the importance of centrality of a node in a network based on the incoming links to that node. A node is given a higher pagerank score if connected to other important nodes. A high hub score means that the node is pointing to many other important nodes. A high authority score means that the node is linked to by many other nodes that are important in the network. Now we can see that the pagerank score for node 1 is 0.009 which is pretty high considering all the scores of other 20 nodes. This means that node 1 is visited the most and hence is central. Now if we see the hub score and authority score of node 1 are very low which indicates that even if the node 1 itself is accessed by important nodes, it doesn't link to important nodes or isn't linked to by many important nodes so we can assume that node 1 is a "bridge" node that connects two otherwise disconnected parts of the network. On the contrary if we see the node 5 we can see that it has a high pagerank and a high hub and authority score which can be justified by the fact that a node with a high pagerank score must be a hub that links to many important nodes or is linked to many other important nodes. Hence, both can be the cases, a node with high pagerank score may have high hub and authority scores or may have low hub and authority scores. Moreover, node 5 had a self loop which may have a positive impact on the pagerank score of a node if the self loop has a high importance in the network which inturn can lead to high hub and authority scores. For node 17 we can see that it has high hub and authority scores but low pagerank score. This is our third and final case where the hub and authority scores are high but the pagerank score is low.