Assignment 3

Mohammad Aflah Khan, 2020082
Neemesh Yadav, 2020529

Dataset Description:

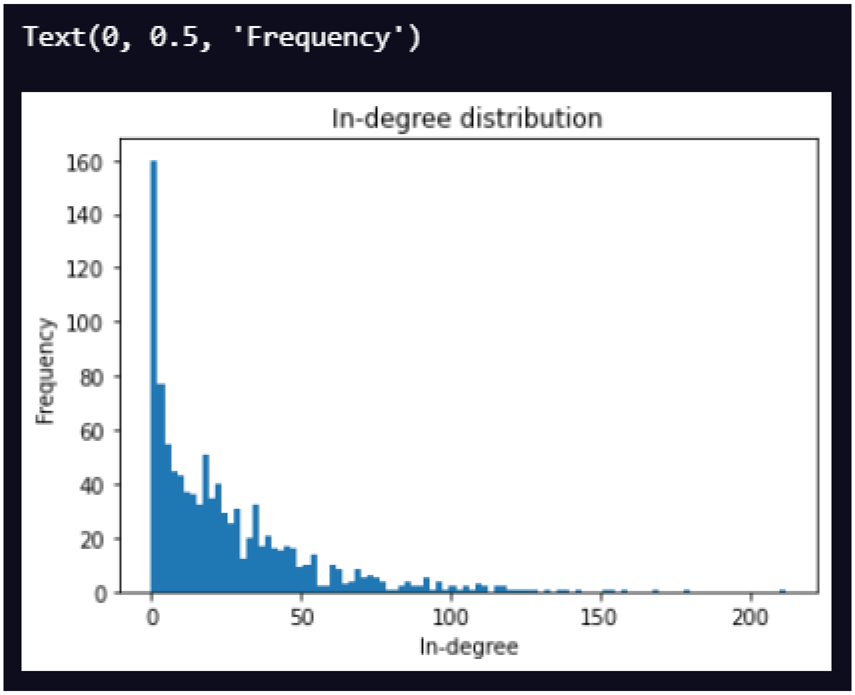
In the dataset, anonymized information about all incoming and outgoing email communication between members of a large European research institution was used to generate a network. The presence of an edge (u, v) in the network indicates that person u sent at least one email to person v. The dataset exclusively represents communication between members of the institution, and does not contain any incoming or outgoing messages to or from individuals outside the institution. Additionally, the dataset includes the "ground-truth" community memberships of each node, where each individual belongs to one of 42 departments at the research institute. The resulting network is a subset of the email-EuAll network, which includes links between institution members and external individuals, albeit with different node IDs.

We chose the “email-Eu-core” dataset for our experiments.

1) Number of unique labels (departments): 42
   Number of nodes: 1005 (Calculated by finding the length of the set of all possible nodes in the dataset, or the columnar length of the adjacency list)
   Number of edges: 25571 (Calculated by finding the sum of the row-wise lengths of each edge in the adjacency list)
   Avg In-Degree: 25.443781094527363 (Calculated by taking average of all the individual node indegrees which are columnar length of the reverse adjacency list)
   Avg Out-Degree: 25.443781094527363 (Calculated by taking average of all the individual node indegrees which are columnar length of the adjacency list)
   Node with Max In-Degree: The node with the maximum indegree is 160 with indegree 212
   Node with Max Out-Degree: The node with the maximum outdegree is 160 with outdegree 334

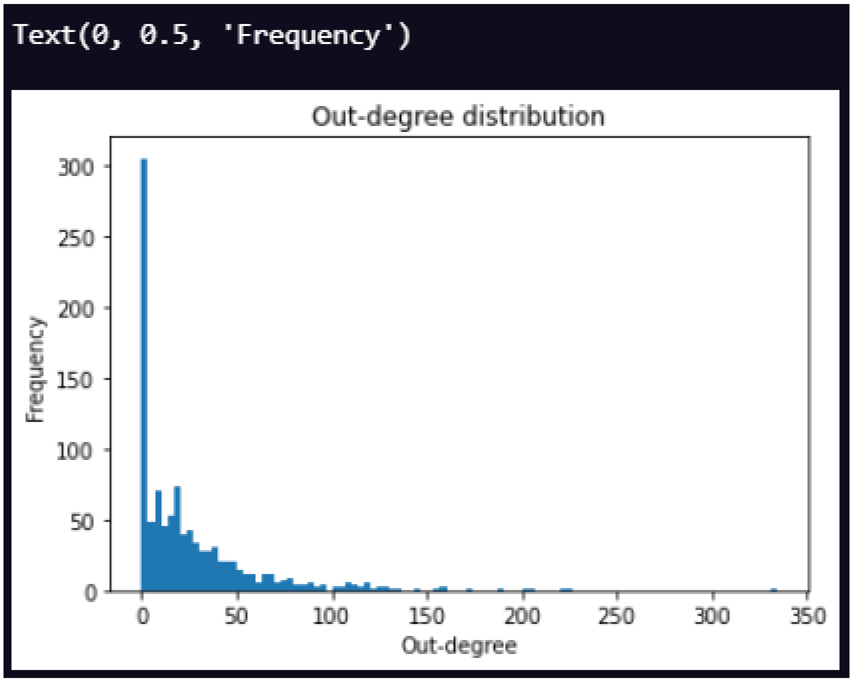
The density of the network: 0.025342411448732432 (num_edges / (num_nodes * (num_nodes - 1)))

a) Degree Distribution for In-Degree:



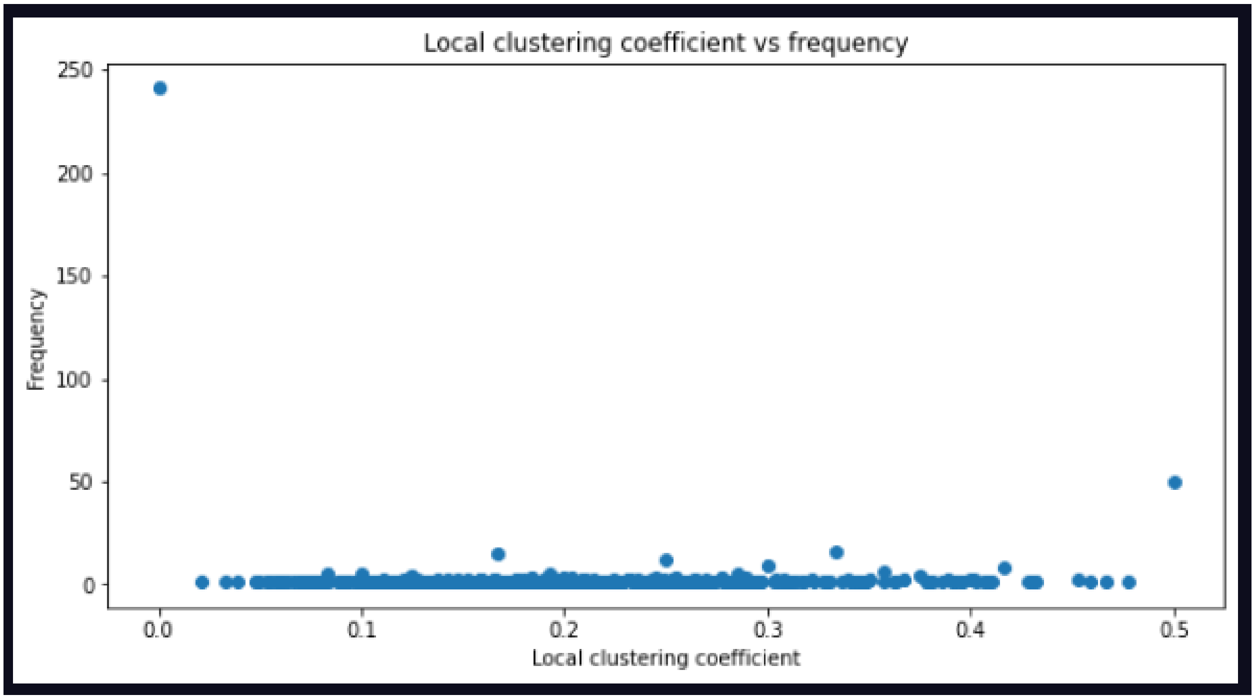
```
Text(0, 0.5, 'Frequency')
```

Degree Distribution for Out-Degree:



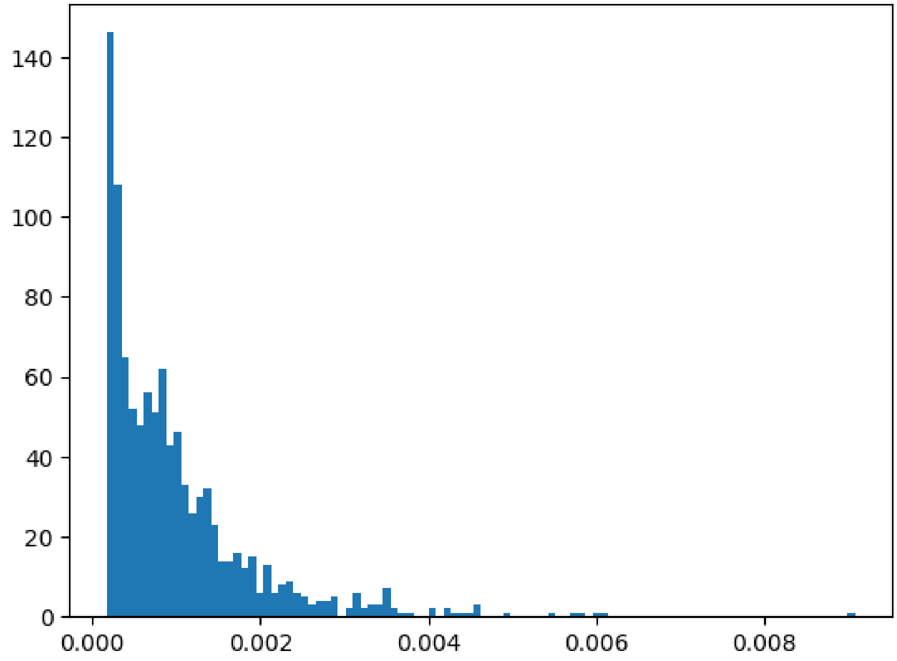
```
Text(0, 0.5, 'Frequency')
```

b) Clustering-Coefficient Distribution:

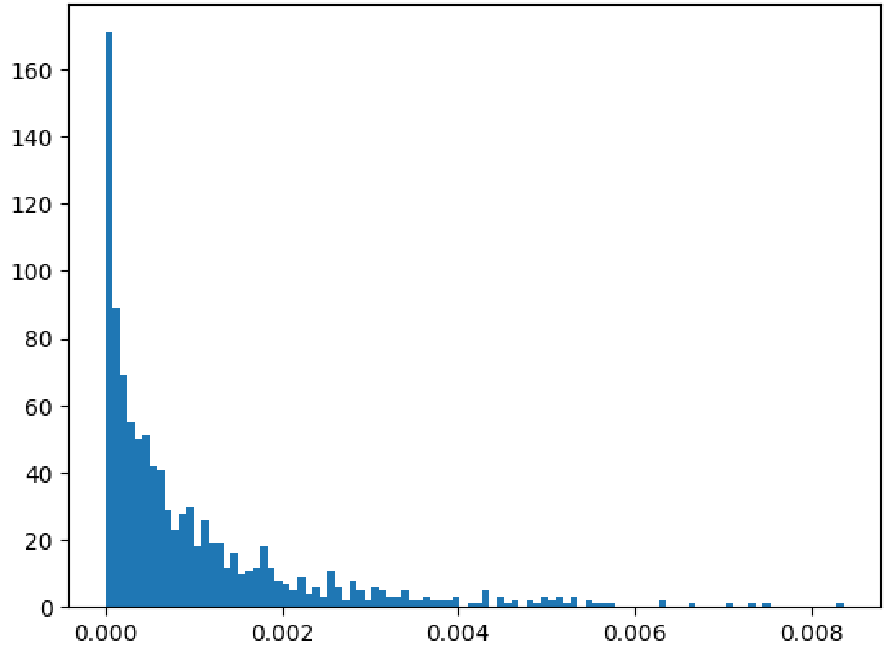


2) We have only shown the distribution/histogram plots of the distributions here, as it is easier to comprehend in that way. The actual values are given in the main.ipynb notebook under the Q_2 directory.

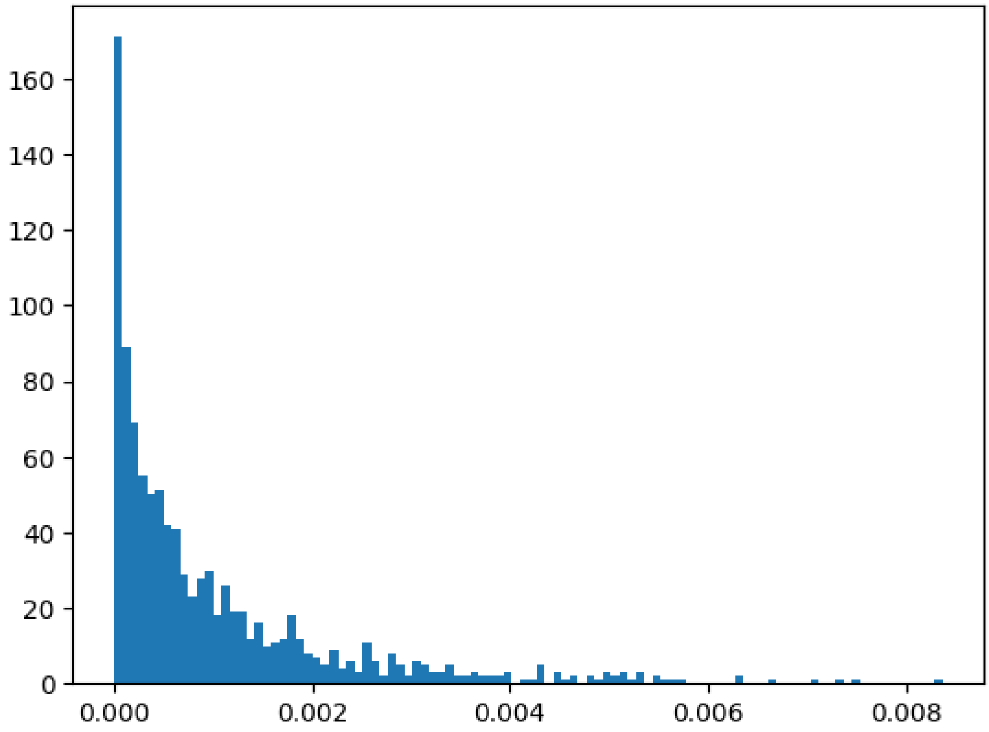
a) PageRank Score Distribution:



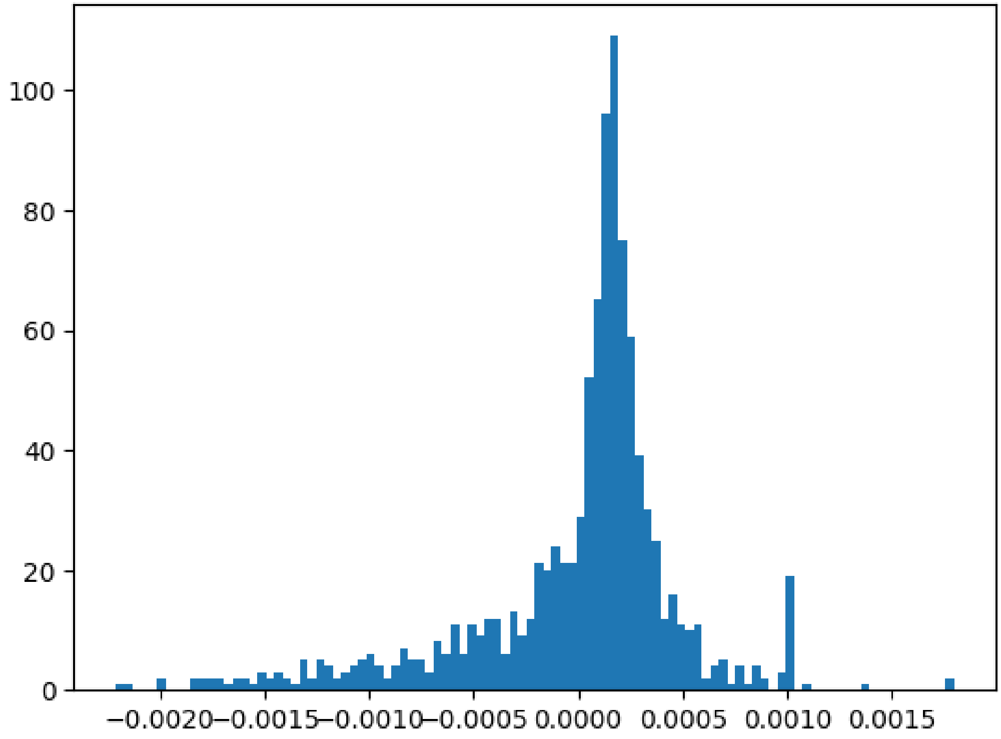
b) Hub Score Distribution:



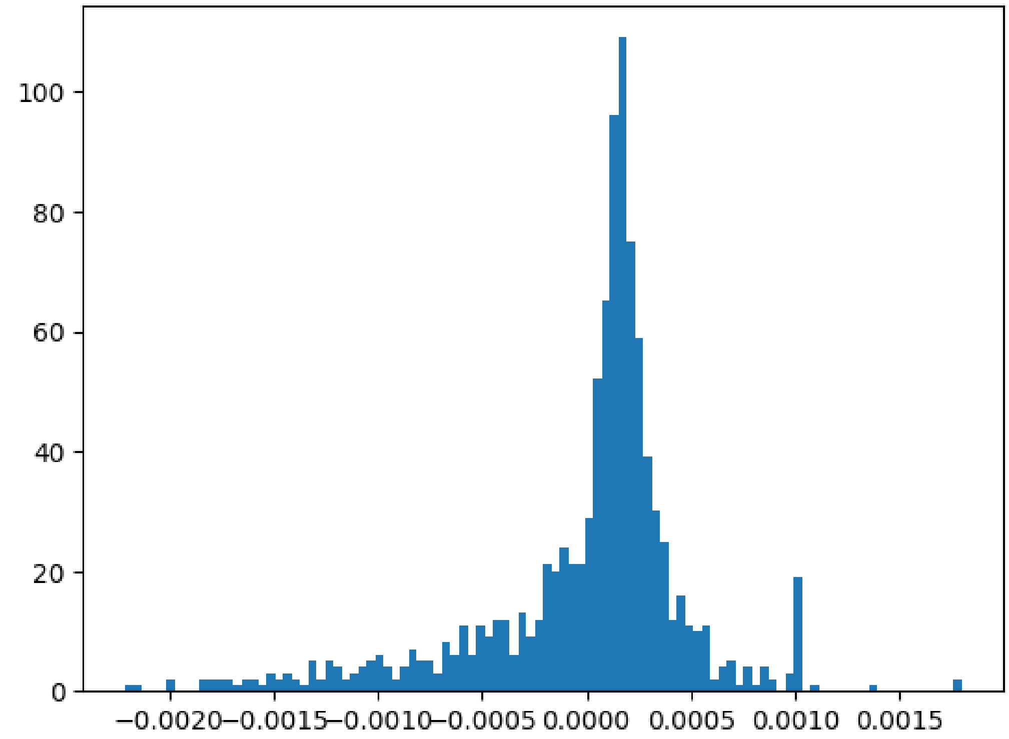
Authority Score Distribution:



c) Difference between the PageRank and Hub Scores:



Difference between the PageRank and Authority Scores:



We see all 3 have similar plot shapes. Authority and Hub are identical while PageRank has a lower peak. This is because incoming and outgoing links are equal and hence the Authority and Hub scores are the same while PageRank also takes into account the structure of incoming links hence there’s a difference between the two.

Pagerank, Authority, and Hub scores are all measures of importance or relevance in different contexts.

Pagerank is a measure of the importance of a webpage in the context of the entire web, based on the number and quality of links pointing to it. It was originally developed by Google as part of its search algorithm and assigns a numerical score to each webpage, with higher scores indicating greater importance.

Authority and Hub scores, on the other hand, are measures of importance within a specific network or community. They were introduced by Jon

Kleinberg in his HITS (Hypertext Induced Topic Selection) algorithm. In HITS, authority scores represent the importance of web pages that are cited or linked to by other important pages, while hub scores represent the importance of web pages that link to other important pages.

In general, Pagerank is a more widely used and recognized measure of importance, particularly in the context of web search. Authority and Hub scores are more focused on specific networks or communities and may be useful for analyzing the structure and dynamics of those networks.

It's also worth noting that while Pagerank, Authority, and Hub scores are all related to the concept of importance or relevance, they may produce different results depending on the specific context and data being analyzed.