Imperial College London

DATA SCIENCE COURSEWORK 3 POSTER (TASK 1)

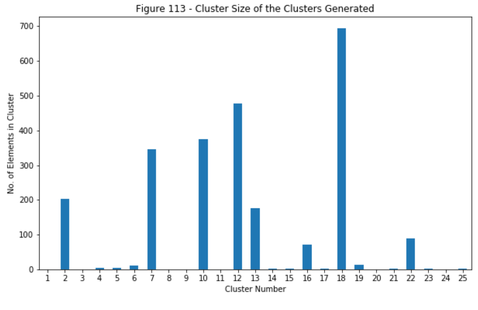
Tudor Trita Trita, CID 01199397

PROBLEM STATEMENT

We are tasked with analysing a collection of scientific papers by keywords. This data set is stored as a feature matrix where each row represents a paper and each column represents a keyword. We also have access to information about citations between the scientific papers. This citation information represents an undirected graph. Our main goal is to extract information about relationships between these scientific papers.

IDENTIFYING CLUSTERS ACROSS THE PAPERS

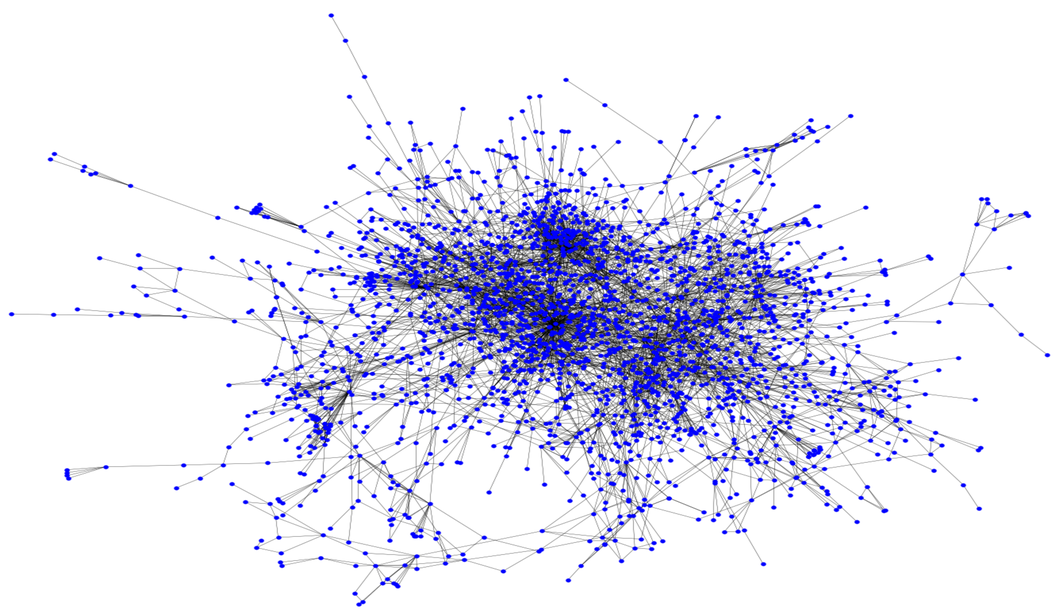
After running the unsupervised clustering algorithm K-Means on our feature matrix, we have found an optimal clustering of 25 clusters.



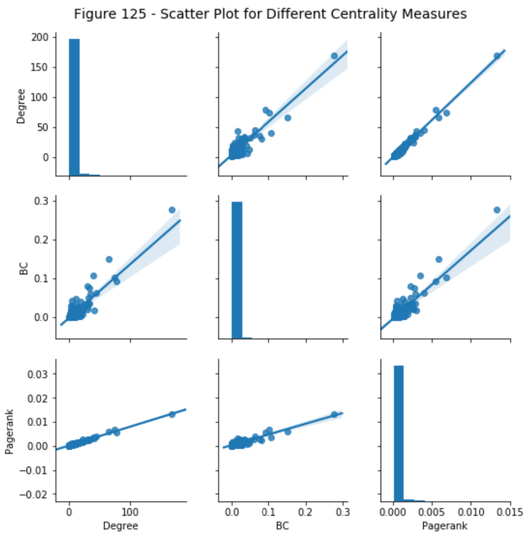
- The bar chart to the left shows the number of elements of each cluster.
- We can see that there are 8 clusters that contain the majority of elements.
- These correspond to the main themes or topics across the scientific papers.
- As we performed unsupervised learning, we don't know what the topics actually are.

ANALYSIS OF CITATIONS BETWEEN PAPERS

We can model the citations between papers as a graph, where each node is a paper cites another paper, we add an edge between them. The resulting graph is below:

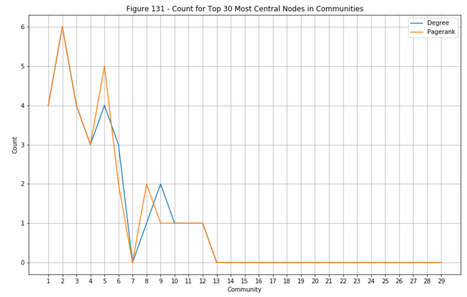


ANALYSIS OF CITATION GRAPH



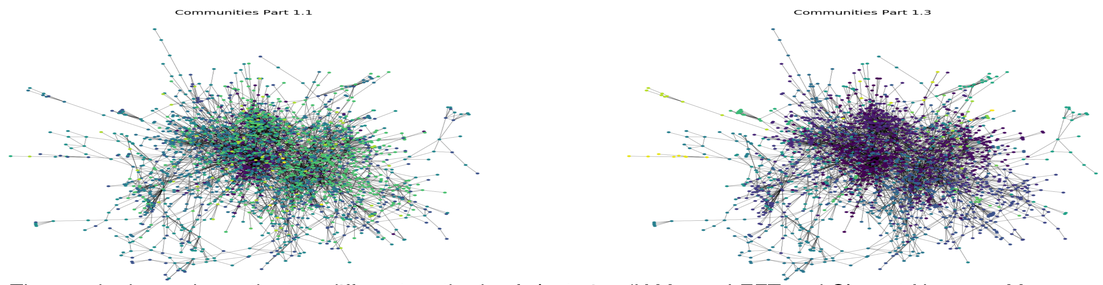
- We considered the three centrality measures for all the nodes in the graph and plotted their relationship.
- We find that the correlations between the measures are strong, indicating agreement.
- The vast majority of the nodes in the graph are not very central, and there are a few hubs with very large centrality measures.
- These hubs represent papers that have been cited extensively, making them very important.
- The 5 most central nodes of the graph are 1245, 271, 1563, 1846, 1672.
- In the next section, we will be showing that these hubs are likely to be the most cited papers in their respective scientific field/category.

DETECTING COMMUNITIES ON THE CITATION GRAPH



- After running the Clauset-Newman-Moore greedy modularity maximisation algorithm, we were able to find the optimal number of communities in the citation graph to be equal to 29.
- We can see that the 30 most central papers (hubs) are located within the first 15 clusters.
- It is not a coincidence that these communities also contain the vast majority of nodes in them.
- We can conclude that each community is composed of a few highly-central nodes within other nodes.

COMPARING BOTH CLUSTERINGS



- The graph above shows the two different methods of clustering (K-Means LEFT and Clauset-Newman-Moore RIGHT) with their respective communities shown in different colours.
- The AUM and ARI index are approximately 0.2 and 0.1 respectively, indicating that these clusterings have little relationship between each other.
- The Clauset-Newman-Moore gives better defined clusters, therefore we recommend to use this algorithm over the K-Means for this use-case.