My Data Science Project

Matthew Cowley

Imperial College London
Department of Mathematics

The Problem

We are looking at how 2485 scientific papers are linked. We do this by using an an Adjacency Matrix 'A' of all the citations of the papers and a Feature Matrix 'F' described below.
A dictionary of size 1433 is created, which indicates presence or absence of a particular key word with in a given paper. This dictionary is used to create a Features Matrix for all the data of size 2485x1433

K-means Algorithm

K-means tries to optimises clusters via inertia. Inertia being the sum of square distance between clustering nodes with the centroids. K-means is done by repeating the steps below for a preset number N- iterations of times and choosing best result w.r.t to Inertia.

1 Selecting a fixed number of nodes to be used as centroids. Are randomly chosen or chosen to be far apart to improve convergence.

2 Clustering are then assigned with optimisation that the inertia in minimised.

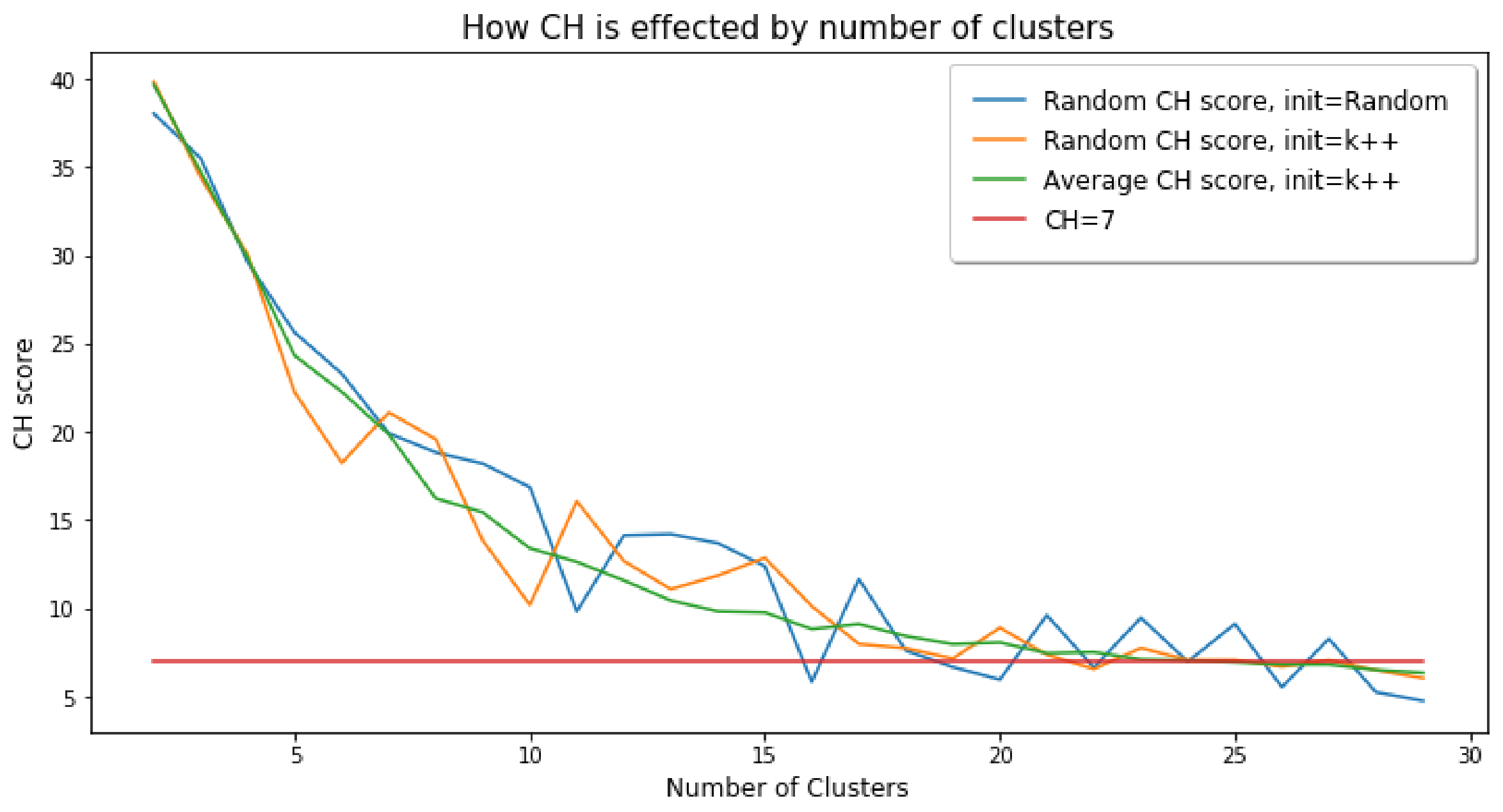
3 Centroids are recalculated.

4 2-3 are repeated until convergence or maximum number of iterations has occurred.

K-means generally gives random results, with the results depending on N-iterations and the data. In this case we use K-means to fit the Feature Matrix, this has very high number of dimensions, so we expected fairly varied results.

Robustness of K-means

When we search for the optimum CH score, we use CH score threshold of '7', we find we get varied results shown by the below graph.



We can see that changing the initial conditions stated in [1] gives similar results, but for 'random' there is far more variation of results for a larger number of clusters. You can also improve the reliability of K-means is by increasing N-iterations.

Clauset-Newman-Moore greedy modularity maximisation (CNM)

The CNN greedy modularity maximisation algorithm optimises modularity of the nodes, the modularity effectively being how many connections there are within a community compared to random number of nodes.

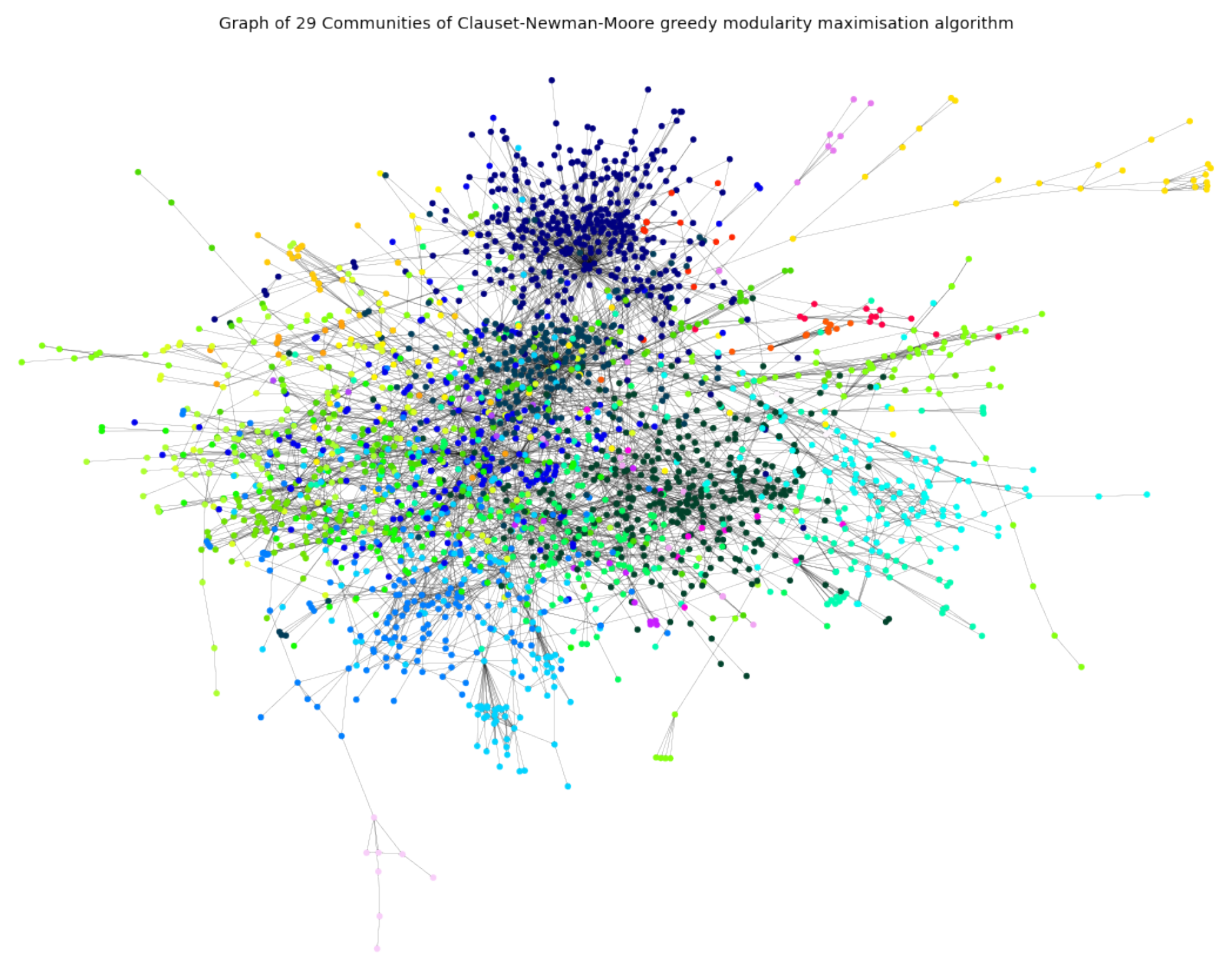


Figure 3: The Communities seem fairly spread out, but do seem to have a dominant cluster in the dark blue colour and light green.

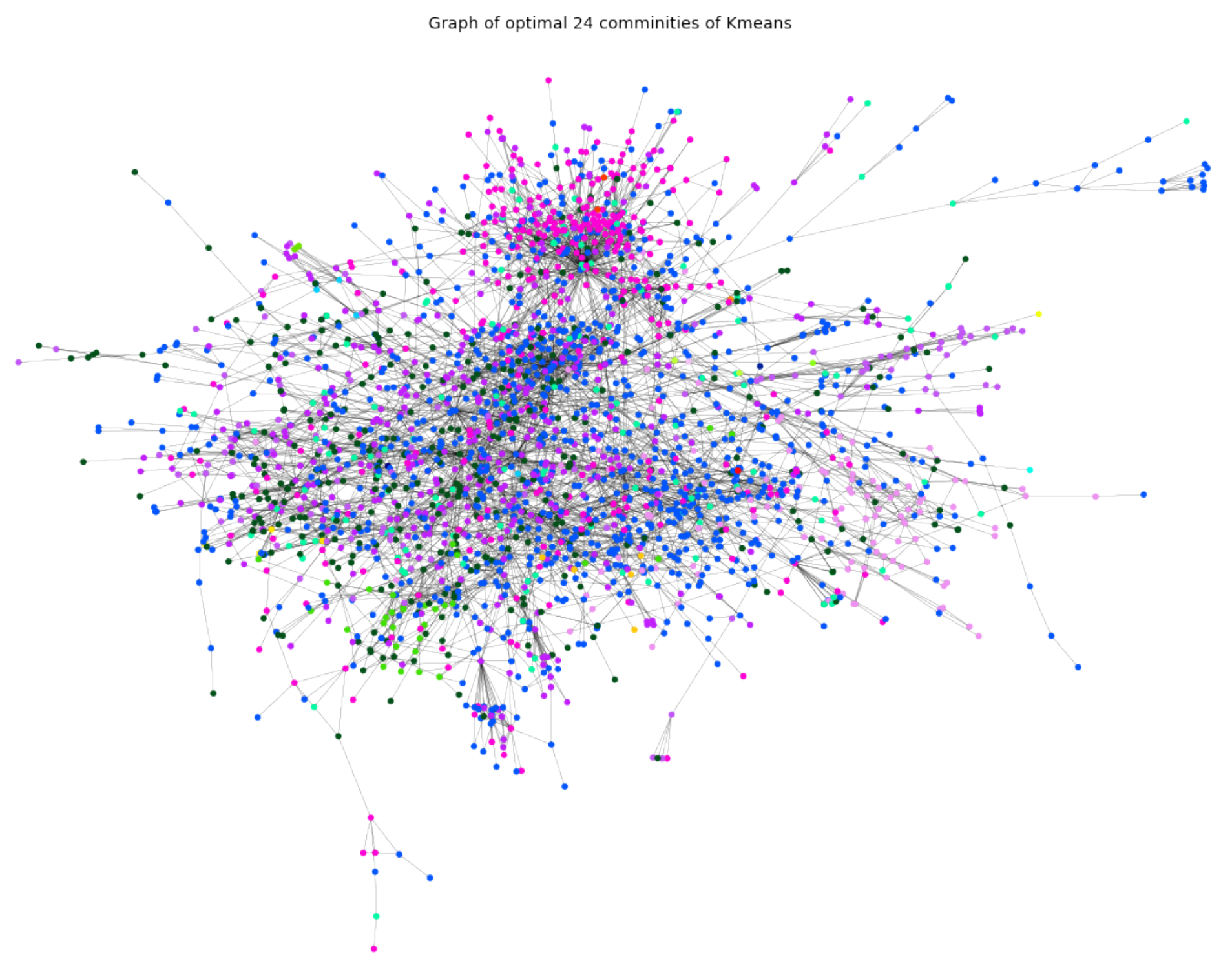


Figure 4: Cluster of K-means for comparison, notice the high overlap and dominance of particular clusterings

Comparison of High Degree Nodes

It is quite interesting to see the behavior of nodes of high centrality as they have the most influence on the clusterings. In this project we looked at 3 centrality's measures: Degree, Pagerank and Betweenness Centrality. As it turns out Degree and Page rank are highly correlated, with just 1 node differing in the the top 30 central nodes respectively.

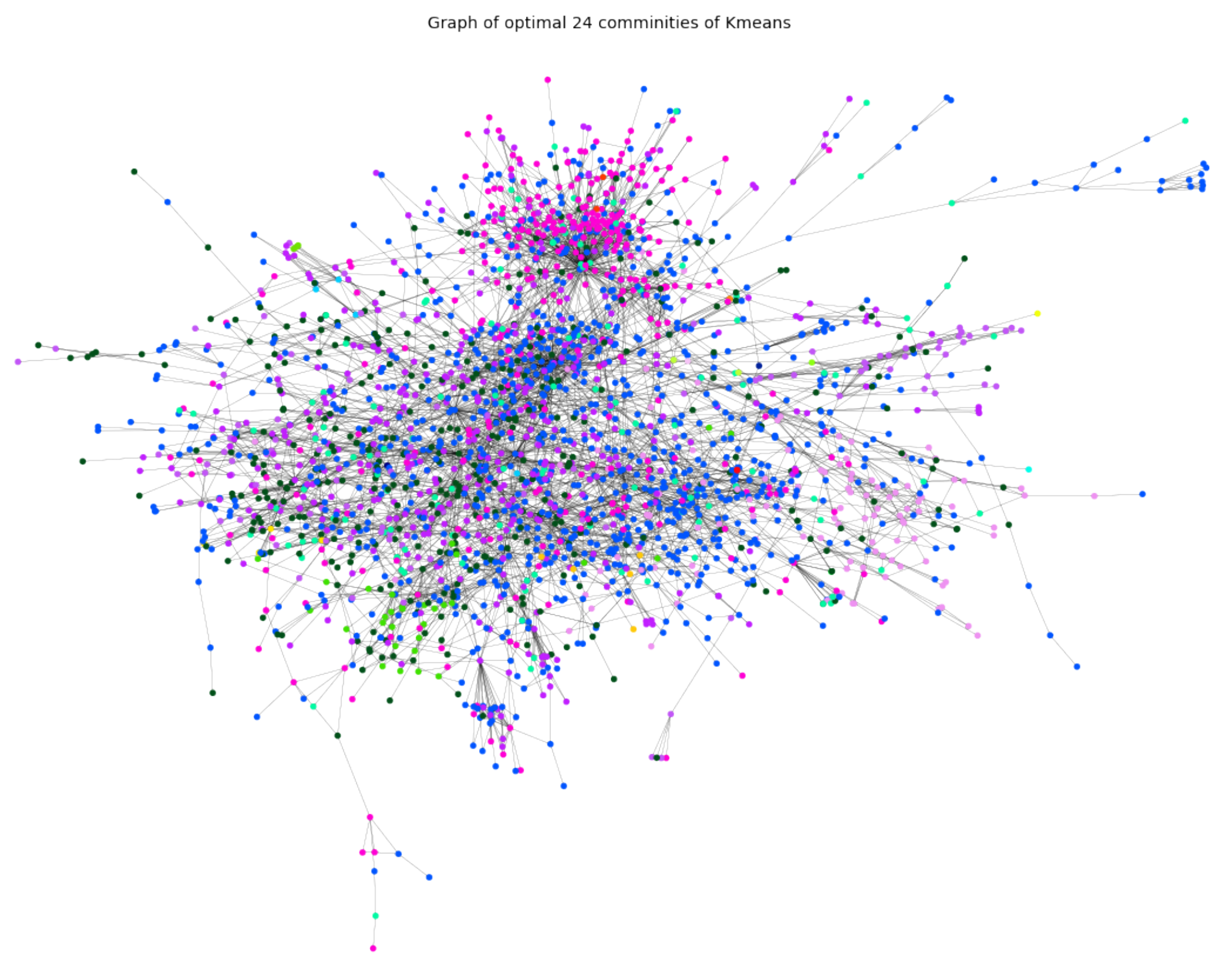


Figure 1: Potential-energy function for particle energy l. [?] As we can K-means has a few clustering which really seem to dominate i.e clustering 7 (blue) in this case. It becomes very apparent how many clustering appear to overlap and this becomes even apparent when you look at all the nodes.

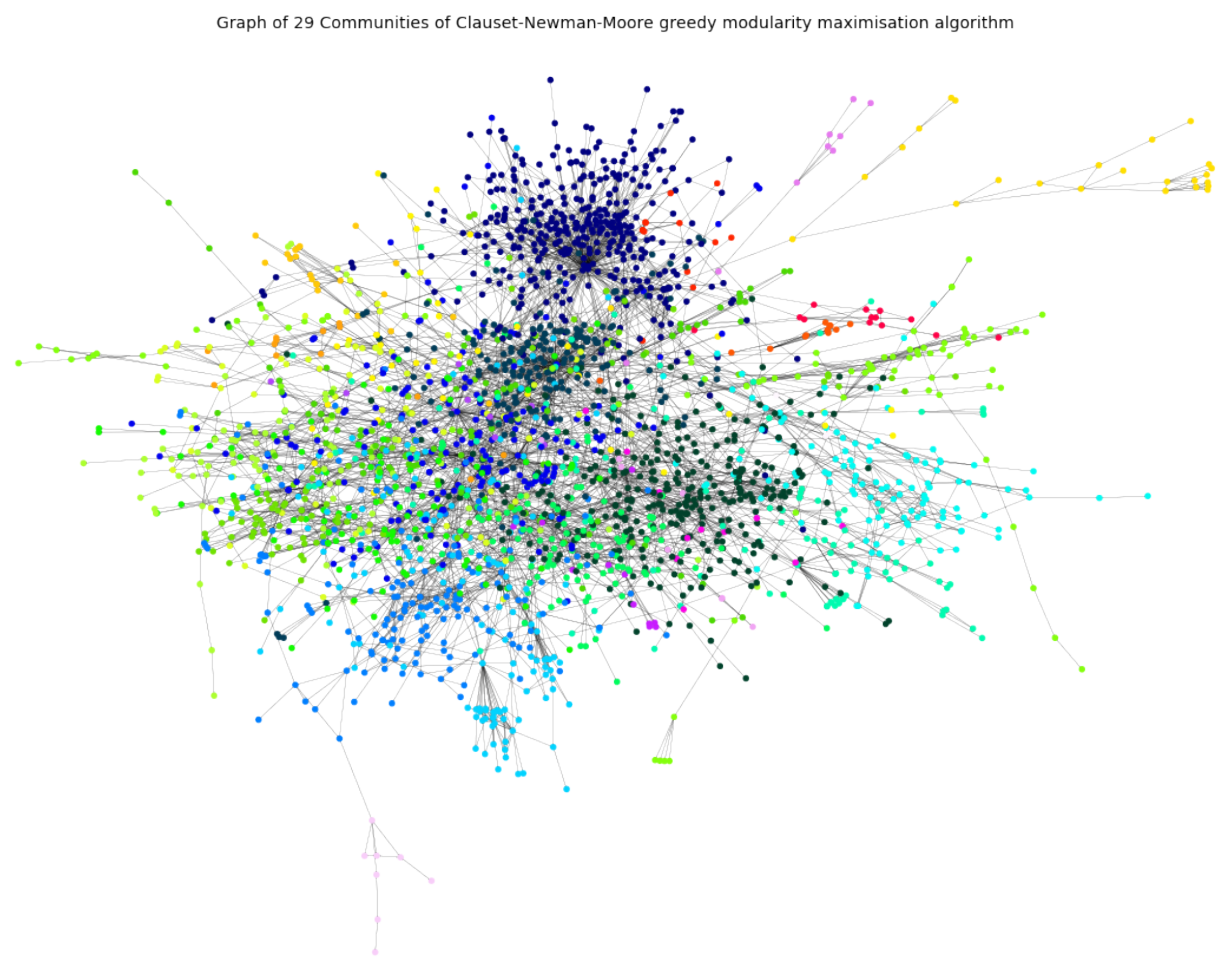
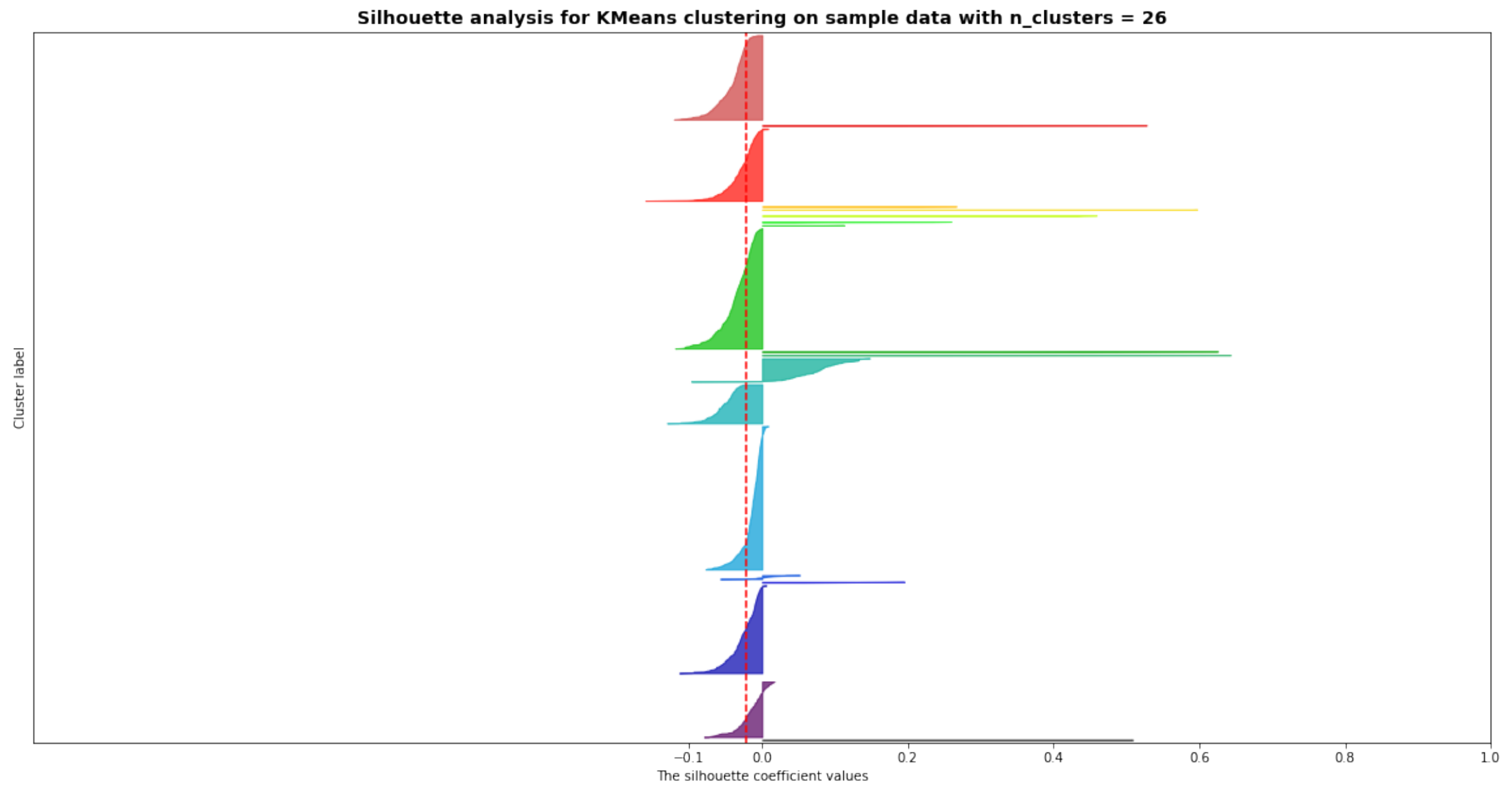


Figure 2: The highly central nodes seem spread out among the communities, which is good sign. There are not very many connections within communities of nodes of high degree apart from communities 0,2 and 5 shown by bold edges.

Silhouette Score of K-Means

The graph below shows the Silhouette score for the Optimum clustering, of 26 clusters.



The red dotted line represents the average score, the score being negative means the clusters are not well defined and so miss labeled. The fact that most scores are near zero suggests that pretty much all clusters are overlapping aswell.

Conclusion

I found that the K-means algorithm performed quite poorly in general. Many indicators pointed towards this e.g the Silhouette score and the Communities graph. It might have had better results if a different method to find optimum clustering was used e.g:

- Finding the optimum clusters is by finding the maximum (or a spike in) in CH score*(k-1).
- Using an elbow method to recognise the optimum.

Even with these changes I believe the model would still have performed poorly due to the high dimensional nature of the problem. The CNM seemed to visually perform better and had far more concrete results with e.g to the Silhouette Score, So in this unsupervised learning problem I would recommended modeling the data as the CNM clusters.

Looking to the Future

It would be good to fit different models e.g K-mediods or potentially do some Hierarchical clustering. It would also be interesting whether better Data or perhaps PCA would improve the results.