PRA302 Assignment 4

Tom Kores Lesjak

September 2024

1 Task 1

1.1 Question 1

Question: Base on its structure, what kind of information do you think could be stored in a graph like this?

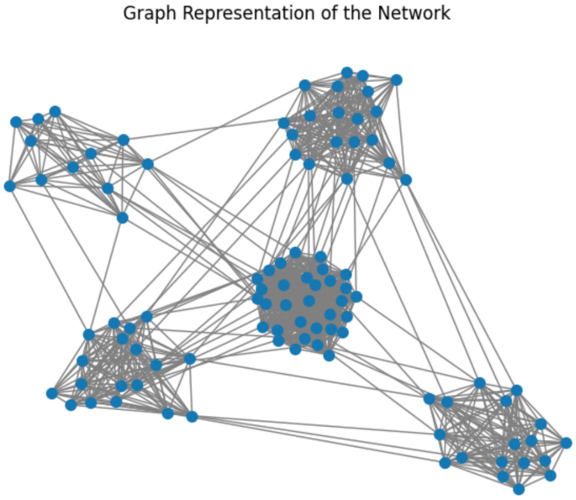

Figure 1: Plot of network given as found in the assignment4_data0 file

The structure of the graph as seen in figure 1 suggests that it could store information about a network of interconnected entities. For example, it could represent model transportation systems, with nodes as locations and edges as routes. It could also represent communication networks where nodes represent devices and edges represent connections between them. Some characteristics of the plot such as the dense connectivity suggests a highly interconnected system, for example applying to supply chain networks, representing the flow of goods between suppliers and distributors.

1.2 Question 2

Question: What is the average path length of the graph?

The average path lenght of the graph is: 2.24

1.3 Question 3

Question: How many communities appear to be present in the graph?

As seen from figure 1, there are 5 communities present in the graph.

1.4 Question 4

Question: What is the intra-community density for each of the communities?

The intra-community density can be found in table 1 and the graph for

Community	**Intra-Community Density**
Community 1	0.9073
Community 2	0.9000
Community 3	0.8824
Community 4	0.9020
Community 5	0.8939

Table 1: Intra-Community Densities for Each Community

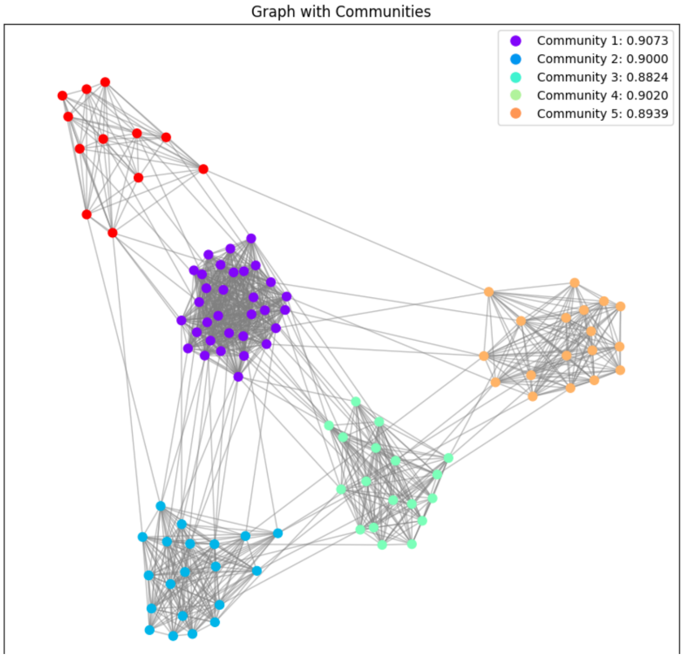

Figure 2: Graph with Communities Colored (Girvan-Newman Algorithm)

1.5 Question 5

Question: What is the inter-community density for the graph?

The inter-community density for the graph is 0.0121.

2 Task 2

2.1 Question 1

Question: Is your implementation of Heymann Garcia-Molina deterministic or stochastic?

The implementation of the Heymann Garcia-Molina algorithm is deterministic. The algorithm creates tag vectors for each taxonomic tag. These vectors are then compared using cosine similarity, which is a deterministic function that consistently computes the similarity between two vectors based on their orientation in a vector space. This ensures that the similarity calculations are the same for each pair of tags when the same data is used. In constructing the similarity graph, an edge is drawn between tags if their similarity exceeds a predefined threshold (α). Since this threshold is fixed and does not involve any random choice, the creation of edges in the graph is deterministic. Also, the closeness centrality of the vertices in the graph is computed. The tag with the highest centrality is selected as the root, based on fixed criteria. When adding remaining tags to the taxonomy, the algorithm uses the cosine similarity values to attach each tag as a child of the most similar tag already in the taxonomy. Since the pairwise similarities are computed once and are fixed, the order in which the tags are added and the structure of the resulting taxonomy will always be the same for the same data.

2.2 Question 2

Question: How many tags are there in each level of your induced taxonomy?

The taxonomy has 1 tag at level 0 and 12 tags at level 1. There are no additional levels in this induced taxonomy.

2.3 Question 3

Question: Explain the relationship between α and the obtained F1 results. Investigate the definition of the F1 score to answer this question.

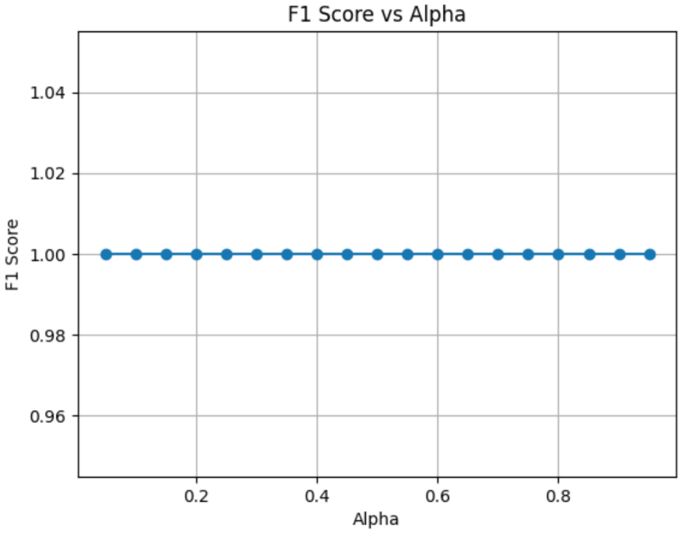

Figure 3: F1 Score vs Alpha values for carnivora data

In our taxonomy induction algorithm, α is the threshold used to determine whether an edge should be added between two tags based on their cosine similarity. Cosine similarity measures how similar two tags are, based on the number of documents they co-annotate. Specifically, if the cosine similarity between two tags is greater than α, an edge is added between the tags. As α increases, fewer edges are added to the graph, as only the most similar tags remain connected. Decreasing α results in more edges, connecting tags with lower similarity. The F1 score is the mean of precision and recall. Precision and recall are defined as follows:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

The F1 score is given by:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{1}$$

A high F1 score indicates a good balance between precision and recall. If either precision or recall is low, the F1 score will also be low, showing imbalances in the model's predictions. The α value directly impacts the number of edges included in the similarity graph, which in turn affects the precision and recall of the induced taxonomy compared to the ground truth. When α is low (closer to 0), more edges are included in the graph. This leads to a decrease in precision, as many false positives are likely to be included in the induced taxonomy. Yet, recall may increase as more true positives are captured. Despite the increase in recall, the rise in false positives typically outweighs the gain and results in a lower F1 score. On the other hand, when α is high (closer to 1) fewer edges are included in the graph. This improves precision as the edges that remain are more likely to be correct. However, recall decreases because many true positives are missed leading to more false negatives. The F1 score decreases when recall is too low, even though precision may be high.

The F1 score is maximized when there is a good balance between precision and recall, which occurs when α allows for the inclusion of meaningful tag relationships without overloading the graph with irrelevant edges. The optimal α is where the trade-off between precision and recall is balanced. The constant F1 score of 0.3097 across all α values, as seen in figure 3, indicates that either the algorithm is not sensitive to changes in α or the dataset itself leads to invariant behavior. Possible issues with the code include uniform cosine similarity values, incorrect graph construction, or errors in the F1 score calculation. However, it is also possible that the dataset's characteristics inherently result in consistent F1 scores.