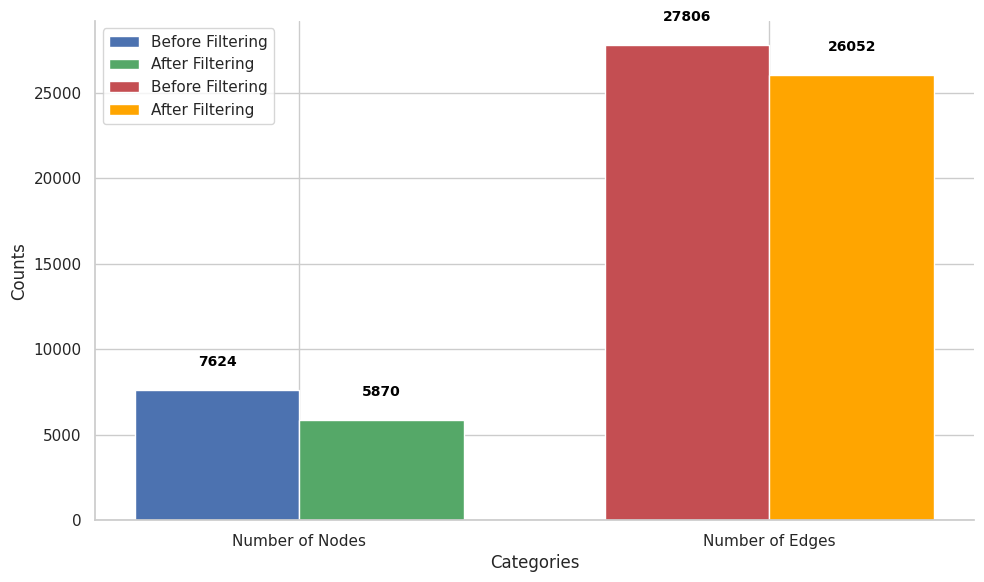
**Network Analysis – Assignment 1**

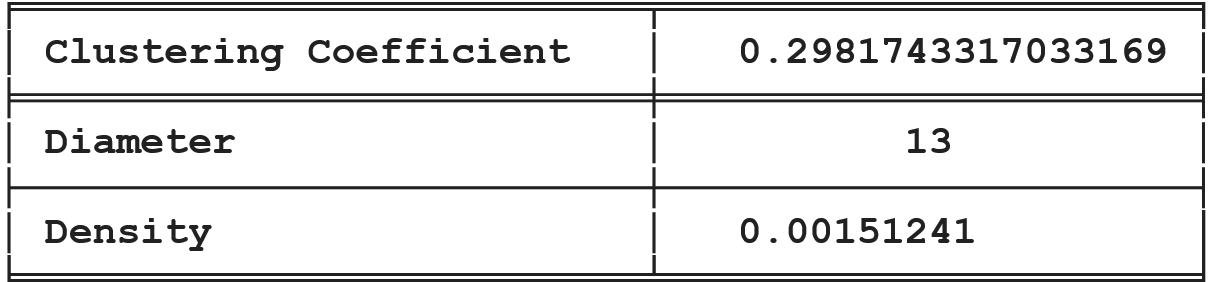
Talal Hussien 208263350 – Daniel Azenkot 318282332

1. The figure below is a bar chart comparing the number of nodes and edges in a graph before and after a filtering process:



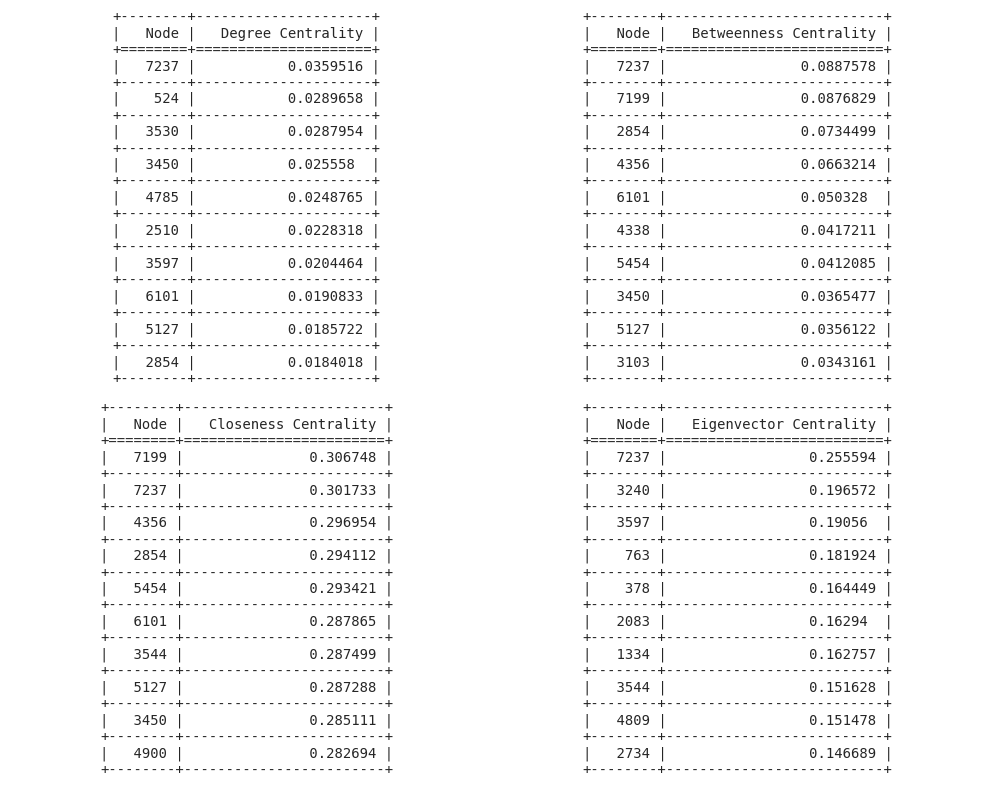
According to the values in the figure, we can see that there were no vertices with a degree of 0, as the difference in the number of edges before and after filtering is equal to the difference in the number of vertices before and after filtering. Additionally, about 23% of the vertices in the graph had a degree of 1, and we will not refer to them in the rest of this work.

1. Main Graph Characteristics :



Based on the clustering coefficient, it can be concluded that, on average, about 29.8% of the vertices in the graph are neighbors of each other. The graph's diameter indicates that the longest path between any two vertices is 13. Additionally, only about 0.15% of the total possible edges in the graph exist. This suggests that there is a medium-sized group of users who are well-connected to each other, while many users have very few connections, as evidenced by the low density and relatively high clustering.

1. In the figure below, we can see tables displaying the top 10 nodes according to each centrality methods

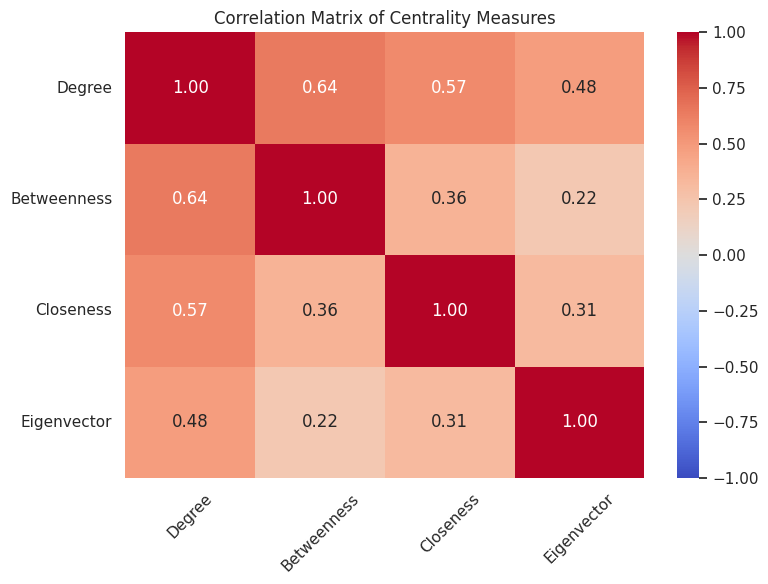


The meaning of each centrality index is as follows: Degree indicates that the higher the degree, the more popular and influential the user is. Betweenness quantifies how much a vertex serves as a bridge between other vertices, specifically whether it lies on the shortest path connecting two vertices. In our network, a user with a high betweenness index acts as a gatekeeper who can potentially control or facilitate the flow of information about artists and songs between different users, because they are positioned between different parts of the graph.

Closeness measures how close a vertex is to all other vertices in the graph, calculated by inverting the average shortest path from it to others. In our network, a higher closeness score means the vertex can more quickly and easily share information about artists or songs it likes, as it is centrally located and can spread information efficiently.

Lastly, the eigenvector centrality index measures how well a vertex is connected within the network, considering connections to well-connected vertices. In our network, users with a high eigenvector centrality are not only well connected themselves but are also connected to other well-connected users, potentially serving as hubs of information for the network.

1. The correlation matrix of the various indices:

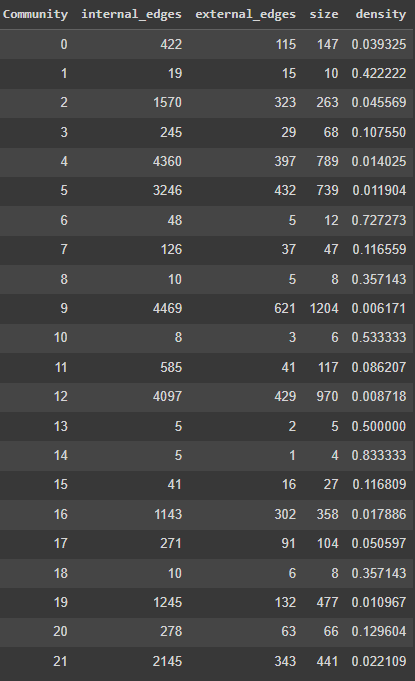


Typically, we expect the measures of degree, closeness, and betweenness to correlate with each other. However, it should be noted that the correlation between closeness and betweenness indices is relatively low at 0.36. Moreover, as demonstrated in the previous exercise, closeness tends to be higher than betweenness, at least among the top 10 vertices.

It may be possible to conclude, as we learned in class (lecture 2 slide 62), that there are vertices in the graph that are close to many others but do not necessarily act as bridges for transferring information. In our network, vertices with a relatively high closeness index—meaning they are close to many other vertices—might not exert as much influence in controlling the flow of information, such as blocking the dissemination of less favored artists, because there are alternative routes through other vertices that do not necessarily pass through them.

1. Yes, different communities can be identified in the graph. We applied the modularity algorithm in Gephi, as learned in class, and found 22 distinct communities with a high modularity score of 0.81.

Table of basic statistics for these communities:



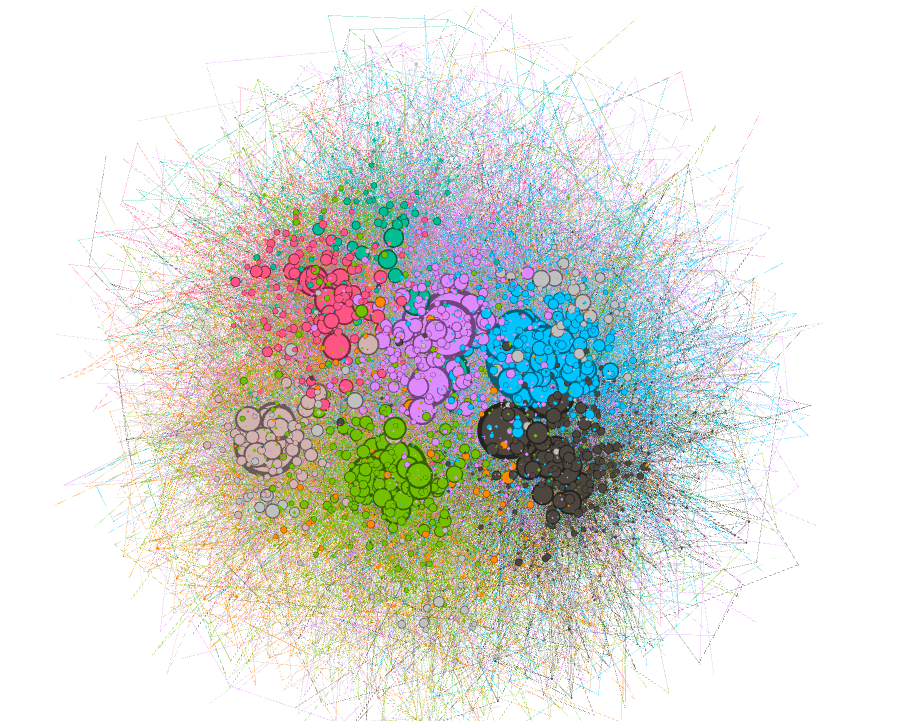
In addition, to present whether there is a connection between the communities created and the countries or artists, we created two normalized frequency tables:

|  |  |
| --- | --- |
|  | A chart of a number of colors  Description automatically generated with medium confidence |

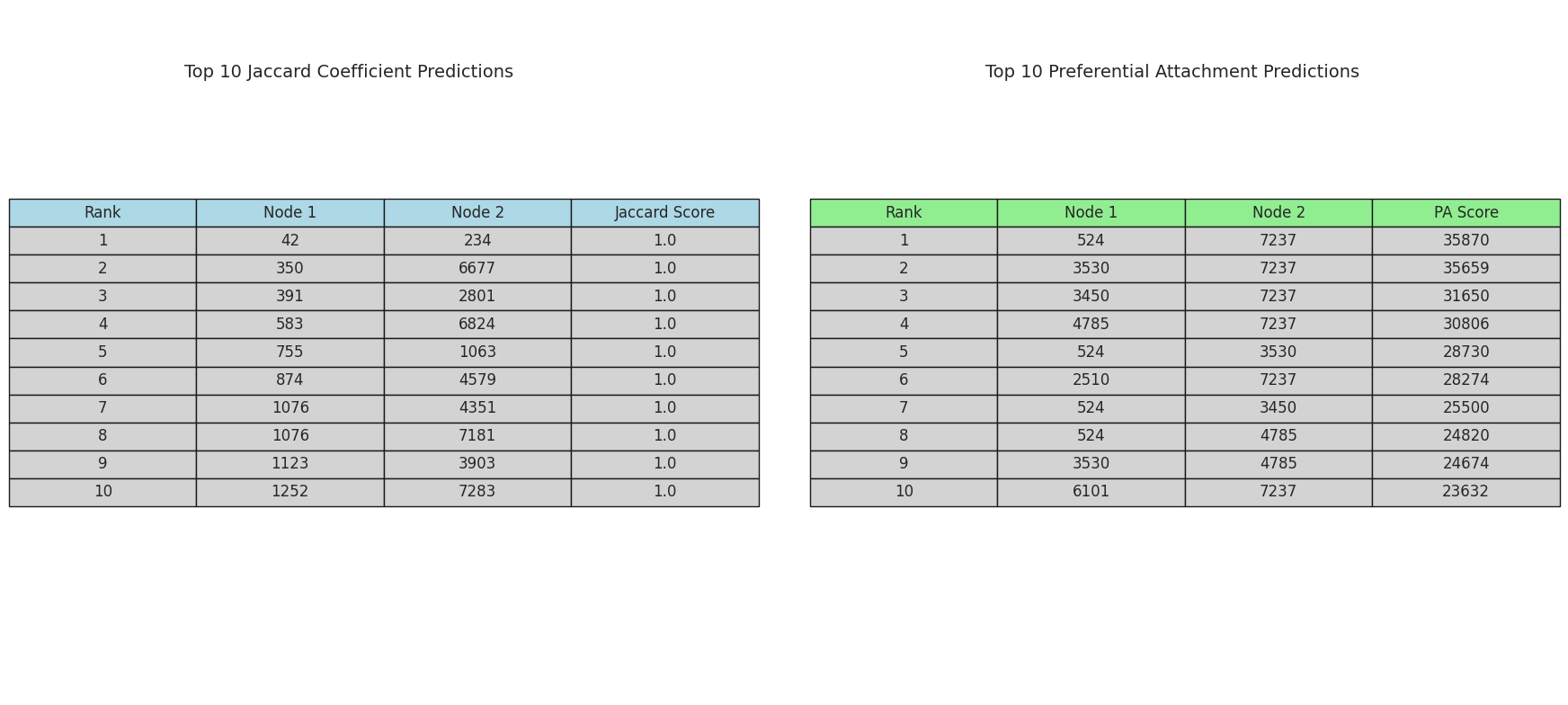
**Left Heatmap :** The country-modularity heatmap suggests that certain countries are more strongly associated with specific modularity classes than others. For instance, countries 1, 3, and 12 exhibit a higher frequency in modularity class 19 compared to other modularity classes. Conversely, there are countries such as 4, 9, and 16 that show low frequencies across all modularity classes.

**Right Heatmap:** The heatmap shows the normalized frequency of top 20 artists across different modularity classes. We can see that in certain modularity classes, almost all the artists have a higher frequency compared to other modularity classes. For example, this is evident in modularity classes 9 and 12. Additionally, we observe that in modularity classes 4 and 5, the artists have a medium frequency compared to other modularity classes.

1. From Gephi, we created another graph that allows us to visualize the communities and degree data in a more aesthetically pleasing and clear manner.



In the network, larger vertices indicate higher degrees, while colors represent different communities. From the graph's structure, it's evident that despite having 22 different communities, there are 6 or 7 central communities containing users with a higher number of connections. This observation aligns with the basic statistics presented earlier, which show several communities with very few users that do not prominently feature in the graph.

1. The table below shows the top 10 predictions for Preferential Attachment and Jaccard Coefficient:

**Jaccard Coefficient:** This method predicts links based on the similarity of nodes' neighbors. All the predicted pairs have a Jaccard score of 1.0, meaning they share all their neighbors. This suggests these nodes are part of tightly knit communities or clusters.

**Preferential Attachment (PA):** This method predicts links based on the "popularity" of nodes, assuming well-connected nodes attract more connections. Node 7237 appears in most top predictions, indicating it's a highly connected node (a hub).

**Do the results make sense?**

* **Jaccard:** Yes, finding nodes with perfectly overlapping neighbors is plausible, especially in social networks with strong community structures.
* **PA:** Also, plausible. Hubs are common in real-world networks, and the PA principle aligns with the idea that popular entities tend to become more popular.

**Additional filters for refinement:**

* Removes users with common friends less than 2 in the same community
* Removes nodes that are overly dominant in link prediction results, promoting diversity by limiting their appearance in the top predictions to a set threshold (e.g., 3 occurrences).

1. In the Excel file, we can see that there are 4 nodes from Country 0 and two nodes from Country 17 appearing in the top 10