Project 1: Networks Theory

Juliana Ramayo
Data Engineering
Universidad Politécnica de Yucatán
Ucú, Yucatán, México
2109128@upy.edu.mx

***Abstract*—This report presents a comprehensive network analysis of the Spotify Artist Feature Collaboration Network, a dataset derived from Spotify's extensive data on artists who have charted on its platform as well as those who have featured in these charting songs. Encompassing a subset of 501 highly followed artists linked through over 4,000 collaborations, the study employs various network analysis techniques to uncover the intricate structures and dynamics of musical collaborations.**

The study not only enhances our understanding of the structural and relational dynamics within the network but also offers actionable insights for music industry stakeholders. By identifying central nodes and community structures, this research aids in discovering emerging talents and facilitating strategic planning for future collaborations.

***Index Terms*—Music Collaboration Network, Social Network Analysis, Community Detection, Spotify Data Analysis**

I. Introduction

The Spotify Artist Feature Collaboration Network is a compelling dataset derived from Spotify, one of the largest music streaming services globally, which boasts over 140 million users. This network encapsulates the collaborative interactions among artists whose tracks have not only charted on Spotify but also those who have featured with charting artists. The dataset encompasses approximately 20,000 artists from the Spotify weekly charts and an additional 136,000 artists involved in at least one collaboration, creating a network of over 135,000 musicians linked by more than 300,000 collaborative edges [1]. For this analysis, a refined sample consisting of 501 artists with the most followers and active collaborations was selected, excluding artists without any collaborative links.

This dataset, which has not been used in prior research, is similar to a study conducted by South in 2018 [2]. South's research delved into the dynamics of the music artist collaboration network using a Breadth-First Search approach to capture the largest connected core of the network. Although the completeness of the network remains uncertain, it predominantly includes popular artists and offers a novel perspective on the dynamics of musical collaborations.

Music is a universal cultural phenomenon that has been part of human civilization for centuries, evolving significantly through the ages. In the digital era, technologies like Spotify have revolutionized how music is consumed and analyzed. The Spotify Artist Feature Collaboration Network offers an invaluable resource for understanding the intricate relationships and influences among artists across various genres.

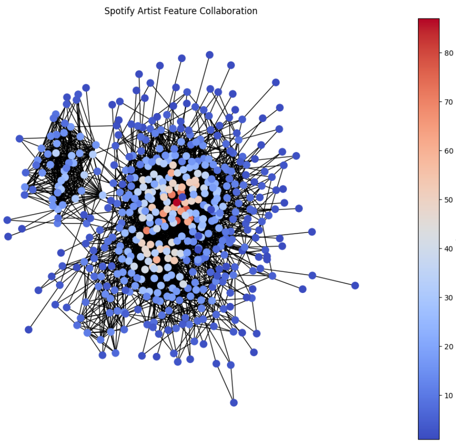

Fig. 1. Spotify Artist Collaboration Network Made with networkx

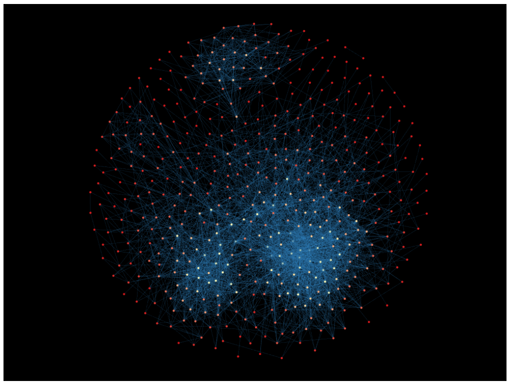

Fig. 2. Spotify Artist Collaboration Network Made with Gephi

The network that will be analyzed is classified as undirected, reflecting the mutual nature of artistic collaborations where each edge represents a bi-directional relationship between artists. This classification underscores the network's function as a social network, where connections imply a shared or collaborative effort rather than a directional flow of influence.

Analyzing this network provides insights into the centrality of artists, the structure of music industry relationships, and the evolution of musical trends, making it a potent tool for stakeholders in the music industry to strategize marketing, discover emerging artists, and predict future collaborations and musical directions.

II. Network Characteristics

The analysis of network characteristics provides fundamental insights into the structure and dynamics of collaborations among top-followed artists. These metrics help us understand the intricacy of the network, identify key players, and discern patterns of connectivity that might influence information flow and collaboration trends in the music industry. To calculate some metrics, the largest connected component was used, which ensures that all considered nodes are reachable from any other node within this component, thereby allowing meaningful computation of distances and paths. For example, in a disjoint graph, many vertices would have infinite eccentricity, rendering several network metrics impractical or undefined. Thus, focusing on the largest connected component provides a more accurate and interpretable analysis of the network's properties [3].

A. Network Size and Number of Links

The network under study includes a total of 501 nodes, with 4186 links connecting these nodes. When considering the largest connected component of the network, which includes 496 nodes connected by 4183 links, we focus on the most interconnected subset of the network.

B. Average Path Length and Distance

The average path length in the network is approximately 3.0840, indicating that on average, any artist is about three steps away from another artist within the largest connected component. This metric shows an "small-world" nature behavior, suggesting that the music industry features close-knit collaborations among top artists.

C. Clustering Coefficient

The clustering coefficient measures the degree to which nodes in the network tend to cluster together. For this network, the clustering coefficient was calculated as 0.317 in Gephi and 0.289 in NetworkX. This variation can arise due to differences in the algorithms or precision levels used by the different tools.

Nonetheless, the result of both tools gives a clustering coefficient of around 0.3, which indicates a moderate level of clustering. This number is a result typical for social networks, where there is a tendency for individuals (or in this case, artists) to cluster into groups or communities. This result shows that there are certain groups of artists who tend to collaborate more frequently among themselves. This could be based on genre, location, or previous collaborations.

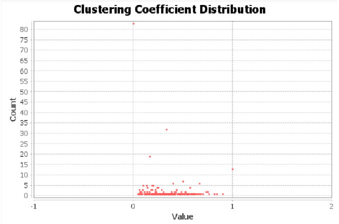

Fig. 3. Clustering Coefficient Distribution by Gephi

D. Diameter

To calculate the diameter, the largest connected was used, since with the original network is not connected which causes the path length to be infinite. The diameter is 8 in both Gephi and NetworkX. The diameter indicates that the furthest distance between any two connected artists is 8 steps.

E. Eccentricity

The eccentricity of a vertex in this connected graph is a critical measure, identifying the furthest distance to any other vertex within the same component. It provides insight into how integral or peripheral an artist is within the network [3].

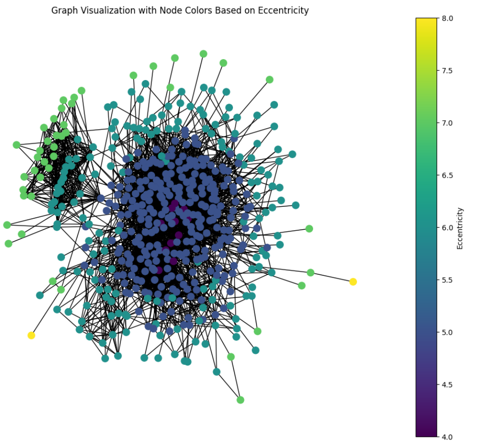

Fig. 4. Graph with Node Colors Based on Eccentricity

F. Radius

There is a discrepancy in the reported radius between Gephi (1) and NetworkX (4). This discrepancy could be due to differences in algorithm implementations or settings within each software. Maybe Gephi is considering the subgraph with only one edge.

The radius measures the minimum eccentricity among all nodes, indicating how closely connected the central nodes are to all others in the network. Taking into account the result from NetworkX, that takes the largest connected component, a radius of 4 means that cenrally located nodes in the network have a maximum distance of 4 steps to the fursthest node thay can reach.

G. Periphery

The periphery of a graph is the subgraph induced by vertices that have graph eccentricities equal to the graph diameter [4], which in this case is 8. The periphery nodes are "Jul" and "NCT DREAM", which means they are the artists that are furthest away from others in the network, and are likely niche or emerging artists.

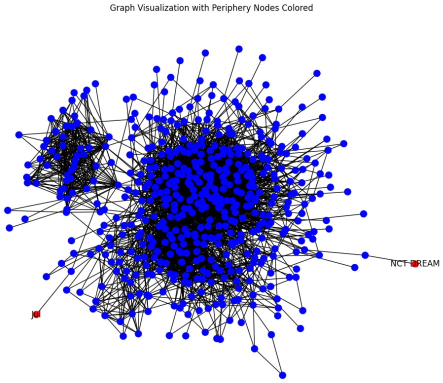

Fig. 5. Graph Highlighting Periphery Nodes

H. Center

The center of a graph is the set of vertices of graph eccentricity equal to the graph radius [5]. In this case, the center nodes were calculated with NetworkX, meaning the center nodes are the ones with eccentricity equal 4.

The center artists in the network are ”Nicki Minaj”, ”French Montana”, ”2 Chainz”, ”Pitbull”, ”Snoop Dogg”, ”Tyga”, ”Maluma”, ”De La Ghetto”, ”Farruko”, ”Ty Dolla $ign”, and ”Flo Rida”. These artists occupy a pivotal role due to their numerous collaborations and due to their strategic positions within the network.

As central nodes, these artists have the ability to influence. This influence can manifest through the spread of musical styles, the introduction of trends, or the facilitation of collaborative projects across genres and regions. Their lower eccentricity means that they act as hubs within the network, enabling them to effectively bridge diverse musical communities and genres.

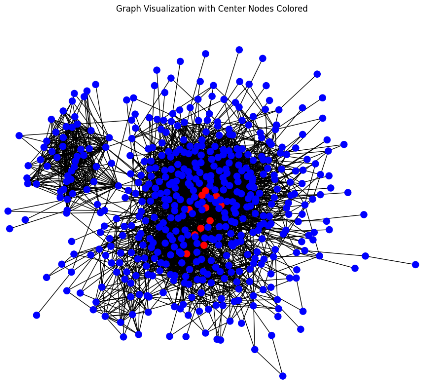

Fig. 6. Graph Highlighting Center Nodes

III. Centrality Measures

A. Degree Centrality

In this network, artists with high degree centrality are those who have collaborated with a large number of other artists.

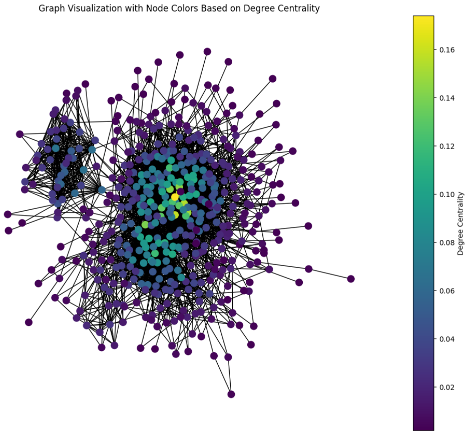
Graph Visualization with Node Colors Based on Degree Centrality
Degree Centrality
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02

Fig. 7. Graph with Node Colors Based on Degree Centrality

In this network, the three higher degree centrality values are owned by Ty Dolla $ign with 0.174, French Montana and Lil Wayne with 0.158, and Nicki Minaj with 0.154. This values represent the proportion of possible connections that each node has relative to the total number of nodes in the network. For instance, a degree centrality of 0.174 suggests that Ty Dolla $ign is connected to approximately 17.4% of all other nodes in the network.

The average degree of the network is 0.0334, which means that, on average, each artist collaborates with about 3.34% of other artists from the network. This relatively low average degree compared to the maximum values for top artists suggests that while a few artists have very high connectivity, a large portion of the network is less connected.

B. Eigenvector Centrality

Eigenvector centrality is a measure of the influence of a node within a network. In this case, a node is considered highly influential if it is connected to other nodes that themselves have high scores.

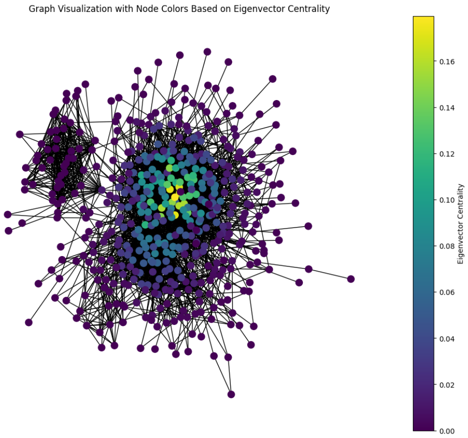
Graph Visualization with Node Colors Based on Eigenvector Centrality
Eigenvector Centrality
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00

Fig. 8. Graph with Node Colors Based on Eigenvector Centrality

In this network, Ty Dolla $ign, French Montana, Future, and Gucci Mane, who exhibit the highest eigenvector centrality scores of 0.1795, 0.1711, and 0.1678 respectively, are not just well-connected but are also connected to other well-connected artists. This implies that these artists are not only active in their collaborations but are also collaborating with other central figures within the music industry.

The average eigenvector centrality of 0.0258 in the network, although significantly lower than the top scores, suggests that while the average artist is somewhat influential, the influence is concentrated among a few top musicians.

C. Closeness Centrality

Closeness centrality is a measure of how close a node is to all other nodes in the network. This centrality measure is particularly useful in understanding how quickly influence can spread from a given node to others across the network.

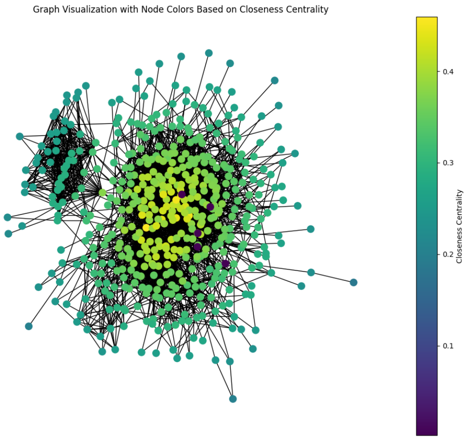

Fig. 9. Graph with Node Colors Based on Closeness Centrality

Artists such as J Balvin (0.4593), Ty Dolla $ign (0.4521), and Snoop Dogg (0.4488) exhibit the highest closeness centrality values. These high values indicate that these artists are, on average, closer to all other nodes in the network compared to other artists. This means that they can spread information, influence, or even trends more efficiently and quickly across the network.

The average closeness centrality of the network is 0.3268, which suggests a moderate level of closeness on average across all nodes. This indicates that while some artists like J Balvin, Ty Dolla $ign, and Snoop Dogg are very well-positioned to disseminate influence, the general artist population in the network has a less central positioning.

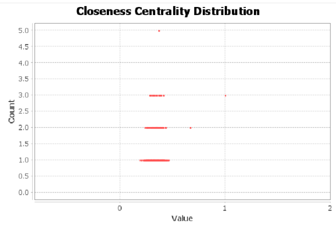

Fig. 10. Closeness Centrality Distribution by Gephi

IV. Degree Distribution

The degree distribution plot shows the number of artists (frequency) against the number of collaborations (degree) they have. This histogram helps in visualizing how connections are distributed among the artists in the network.

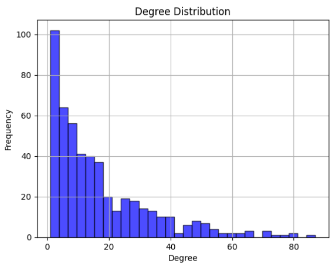

Fig. 11. Degree Distribution Plot

A significant number of artists have relatively few collaborations, as evidenced by the high bars at the lower end of the degree spectrum. This indicates that many artists tend to collaborate with only a small group of other artists or are possibly newer to the scene. Moreover, the plot shows a long tail to the right, which tapers off as the degree increases. This suggests the presence of a few highly connected artists, or "hubs," which have a large number of collaborations.

V. Community Detection

The chosen method for community detection was hierarchical clustering, which is a method that builds communities by progressively merging nodes or groups of nodes based on their similarity. In this context, hierarchical clustering was applied to group artists into four distinct clusters. This approach helps in understanding the interaction and collaboration patterns among artists, possibly reflecting underlying genres, collaboration styles, or other affiliations.

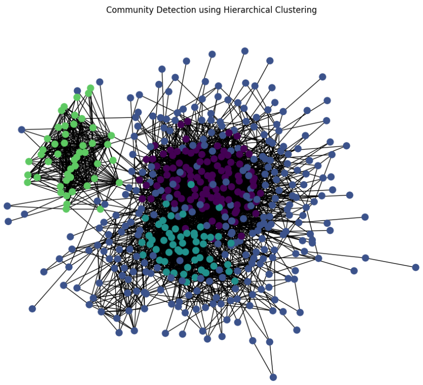

Fig. 12. Community Detection using Hierarchical Clustering

To measure the quality of this communities, modularity was used. Modularity quantifies the strength of division of a network into clusters by comparing the density of links inside communities with the density expected in a randomized network. A higher modularity value indicates a structure that deviates more from randomness, suggesting a meaningful community structure [6].

The modularity of the communities using hierarchical clustering was of 0.41, which indicates a moderate community structure relative to what would be expected in a random network model.

VI. Conclusions

This analysis of the Spotify Artist Feature Collaboration Network through various network analysis techniques has yielded significant insights into the structural dynamics and interaction patterns within the music industry. For example, the use of hierarchical clustering for community detection has proven particularly effective, with a modularity score of 0.41 indicating well-defined community structures. These communities likely represent groups of artists connected by similar genres, collaborative tendencies, or geographical ties.

Moreover, the network structure, revealed through metrics such as degree distribution, centrality measures, and clustering coefficients, underscores the presence of influential hubs and close-knit collaborations that could influence musical trends and marketing strategies.

Understanding these network dynamics offers valuable opportunities for stakeholders in the music industry to optimize their engagement strategies. For instance, record labels and music marketers can tailor their promotional efforts based on the identified central nodes and community structures, ensuring more targeted and effective outreach. Additionally, the network analysis provides strategic insights that can aid in discovering emerging talents and predicting future collaborations.

References

[1] J. Freyberg, “Spotify Artist Feature Collaboration Network,” *Kaggle,* Oct. 10, 2022. https://www.kaggle.com/datasets/jfreyberg/spotify-artist-feature-collaboration-network/data?select=edges.csv (accessed Jun. 21, 2024).

[2] T. South, “Network analysis of the Spotify Artist Collaboration Graph,” *Vacation Research Scholarship 2017-2018,* 2017, [Online]. Available: https://vrs.amsi.org.au/wp-content/uploads/sites/84/2018/04/tobin_south_vrs-report.pdf

[3] Wolfram Research, Inc., “Graph Eccentricity,” *Wolfram MathWorld,* Jun. 18, 2024. https://mathworld.wolfram.com/GraphEccentricity.html (accessed Jun. 22, 2024).

[4] Wolfram Research, Inc., “Graph Periphery,” *Wolfram MathWorld,* Jun. 18, 2024. https://mathworld.wolfram.com/GraphPeriphery.html (accessed Jun. 22, 2024).

[5] Wolfram Research, Inc., “Graph Center,” *Wolfram MathWorld,* Jun. 18, 2024. https://mathworld.wolfram.com/GraphCenter.html (accessed Jun. 22, 2024).

[6] F. Menczer, S. Fortunato, and C. A. Davis, *A first course in network science.* Cambridge University Press, 2020. doi: 10.1017/9781108653947.