Project 3: Link Prediction

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

***Abstract*—This report delves into the analysis of social networks derived from collaboration patterns among artists, focusing on two main networks: the user's top artists and the top artists in Mexico. Employing advanced network analysis techniques, this study identifies prevailing collaboration patterns and predicts potential future collaborations, while also exploring the underlying community structures within these networks. Link prediction methods such as the Jaccard coefficient and resource allocation provide deeper insights into future collaboration possibilities, thus offering valuable guidance for strategic planning in the music industry. This comprehensive approach not only enhances understanding of current musical trends but also aids in anticipating future developments, making it a crucial tool for artists, producers, and industry stakeholders navigating the complex landscape of music production and distribution.**

***Index Terms*—Music Collaboration Network, Music Industry Trends, Link Prediction**

I. Introduction

In the digital age, the creation, distribution, and consumption of music have been profoundly transformed by technological advancements and the rise of streaming platforms like Spotify. This report aims to analyze social networks constructed from collaboration patterns among artists, focusing on two distinct networks: one comprising artists frequently listened to by the user, and another consisting of the top artists in Mexico. The primary goal is to uncover collaboration patterns and potential community structures within these networks, offering insights into how artists interact and how these interactions could influence future collaborations.

The networks analyzed are made up of nodes representing individual artists, with links denoting collaborative ties between them. This approach helps in understanding not just the musical ties but also potential social and professional relationships among the artists. Such analysis is particularly valuable in the music industry, where strategic collaborations can significantly impact an artist's reach and success. Furthermore, this network analysis might reveal broader social dynamics, such as the formation of influential artist clusters or communities.

No specific attributes are associated with the nodes or links for this analysis, keeping the focus on the existence and frequency of collaborations. The networks are dynamic, reflecting real-time data based on current listening trends and the yearly popularity of artists. This real-time aspect means the structure and significance of the networks can shift, reflecting the ever-changing landscape of music popularity and artist collaborations.

II. Mapping Process

To analyze the network of artist collaborations, data was accessed using the Spotify API [1] through the Spotipy library in Python [2]. This approach allowed for the extraction of detailed information about artist collaborations and user listening habits. The Spotify API was primarily used to gather data on top artists based on user listening history and top artists in Mexico, reflecting real-time trends and popular collaborations.

In this network, a link between two nodes (artists) was established based on their collaboration on top songs. Due to the Spotify API's request rate limits, the scope was refined to focus on top songs, representing a pragmatic approach to understanding high-profile collaborations. Furthermore, in the beggining of the project, data collection faced challenges due to limitations in scraping the Spotify Charts website, which required a complex login process. This challenge was circumvented by utilizing the Spotify API, which provided a more straightforward and compliant method to access the necessary data. Selecting specific markets (e.g., Mexico) for top artists helped refine the data collection process to manageable and relevant subsets of artists.

The data extraction process involved key functions implemented in Python with two functionalities:

- **Top Artists Retrieval:** Artists frequently listened to by the user and top artists in Mexico were fetched. This involved handling pagination and fetching details like artist name, popularity, and genres.
- **Collaboration Identification:** The get_collaborations function was used to identify actual collaborations among the top artists by examining their top tracks and identifying common artist appearances on these tracks.

Data was cleaned and processed within these functions, ensuring that only relevant and accurate artist details were retained for network analysis. This process included filtering and validating data to ensure consistency across the extracted datasets. To ensure the network accurately mirrored actual artist interactions, the data was cross-referenced with publicly available collaboration information on Spotify. This validation step was crucial to confirm that the mapped collaborations closely represented the real-world artistic interactions, thereby enhancing the reliability of the network analysis.

For visualizing and analyzing the network, Python libraries such as Matplotlib and NetworkX were employed. These tools facilitated the creation of network graphs that visually represented the relationships and collaborations between artists, providing a clear and insightful depiction of the community structure within the network.

III. Basic Characteristics and Visualization

A. My Top Artists

My network of top artists comprises 101 nodes and 128 edges, reflecting a moderately connected structure typical of collaborative networks in the music industry. The undirected nature of this graph confirms that collaborations are mutual, without hierarchical implications.

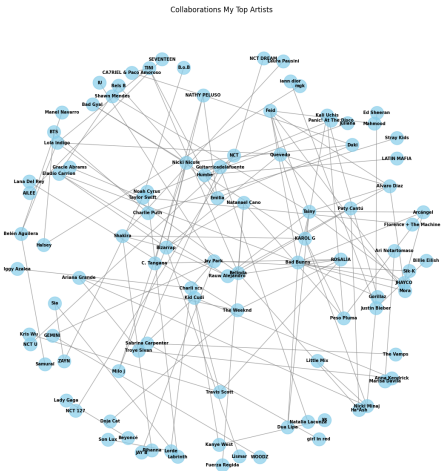

Fig. 1. Collaborations My Top Artists

1) Connectivity: The network features disjointed components, indicating that not all artists within my preferred categories have direct collaborations. This is a common occurrence in networks where musical tastes are diverse and genres or regional influences might not overlap. The largest connected component includes 66 out of the 101 nodes, representing a significant, though not dominant, part of the network.

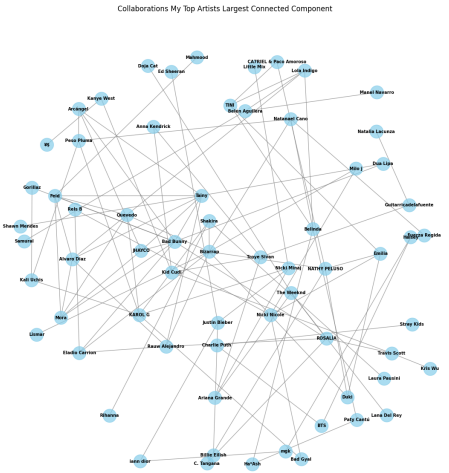

Fig. 2. Collaborations My Top Artists Largest Connected Component

2) Clustering and Path Length: The average clustering coefficient is 0.1119 for the entire network, increasing to 0.1712 within the largest connected component. This low overall clustering suggests that my musical preferences span artists who infrequently collaborate closely, indicating a broad diversity in genres and styles. The average path length of 4.651 in the largest connected component implies a typical degree of separation within larger networks.

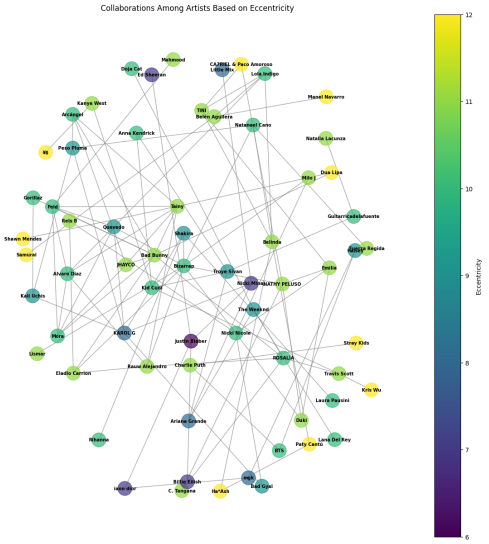

Fig. 3. Collaborations Among Artists Based on Eccentricity

3) Network Diameter and Centrality: The network's diameter of 12 and radius of 6 suggest a wide dispersion among the nodes, with Justin Bieber emerging as the most central node. This indicates his broad collaborative reach within the network. While the periphery features artists like Stray Kids and Shawn Mendes, whose links are comparatively more distant.

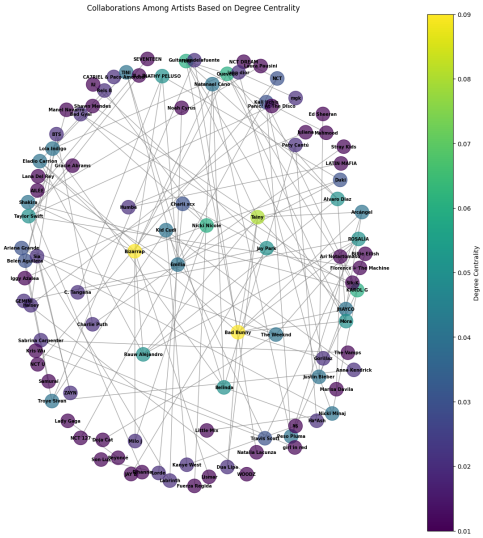

Fig. 4. Collaborations Among Artists Based on Degree Centrality

4) Degree Centrality: Degree centrality highlights the number of direct connections each artist has. Bad Bunny and Bizarrap, each with the highest scores, demonstrate widespread collaborations that significantly influence and connect various musical styles and communities within the network. This reflects my inclination towards artists who introduce me to diverse musical styles.

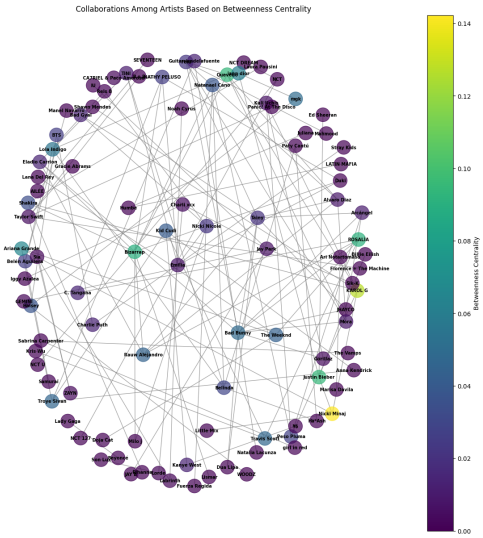

Fig. 5. Collaborations Among Artists Based on Betweenness Centrality

5) Betweenness Centrality: Betweenness centrality further underscores the role of artists like Nicki Minaj, who with the highest score of 0.1427, acts as a crucial bridge within the network. Her collaborations often connect disparate musical styles, emphasizing her role in diversifying my musical exposure.

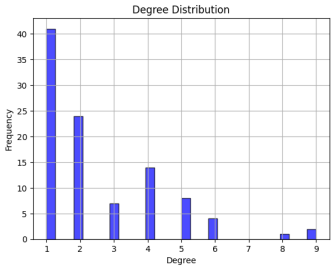

Fig. 6. Degree Distribution

6) Degree Distribution: The degree distribution shows a high frequency of artists with only one or two collaborations, indicating a network rich with diverse, possibly niche artists who do not frequently engage in mainstream collaborations.

This analysis has revealed the structural characteristics of my music listening habits and it also highlights the broad and eclectic range of my musical preferences, driven by artists who are both highly collaborative and those who form unique, less connected nodes within the network.

B. Mexico's Top Artists

The network of Mexico's top artists is substantially larger and more interconnected than the previous network, consisting of 156 nodes and 465 edges. This suggests a vibrant and collaborative music scene in Mexico, characterized by frequent collaborations among artists. The undirected nature of the graph indicates mutual relationships without hierarchical distinctions.

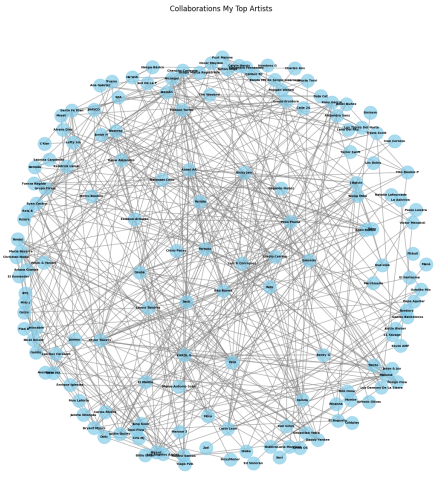

Fig. 7. Collaboration Mexico's Top Artist

1) Connectivity and Clustering: Despite its size, the network is not fully connected, which is typical for large networks where some artists or genres might be less integrated. The largest connected component includes 150 out of 156 nodes, demonstrating a high level of interconnectivity among the majority of artists. The average clustering coefficient of 0.3054 and the clustering coefficient of the largest component at

0.3176 both suggest a moderate level of clustering. This indicates that groups of artists tend to collaborate more intensively within their clusters, possibly around specific music styles or regional affiliations.

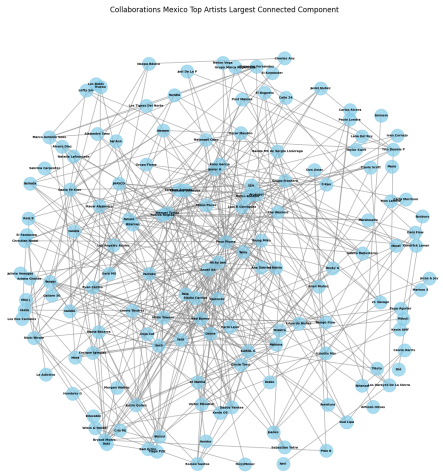

Fig. 8. Collaborations Mexico Top Artists Largest Connected Component

2) Network Metrics: The average path length of 3.767 and a diameter of 10 in the largest connected component indicate that any two artists are, on average, less than four steps away from each other, which facilitates the spread of musical influences and trends quickly across the network. The network's radius of 5 and central node, KAROL G, highlight her as a pivotal figure in connecting various musical genres within the industry.

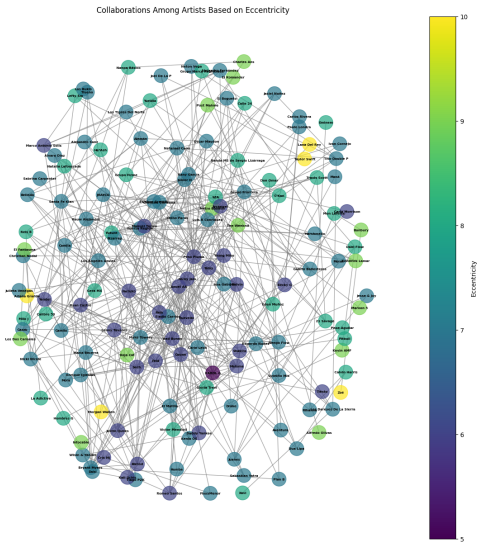

Fig. 9. Collaborations Among Artists Based on Eccentricity

3) Degree Centrality: Peso Pluma and Bad Bunny are among the artists with the highest degree centrality, indicating they have the most direct connections to other artists. This suggests they are central figures in Mexico's music scene, potentially serving as influential nodes through which collaborations and new music trends proliferate.

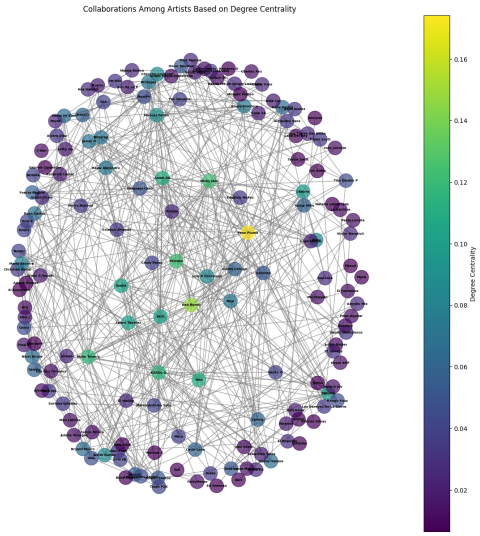

Fig. 10. Collaborations Among Artists Based on Degree Centrality

4) Betweenness Centrality: Peso Pluma also leads in betweenness centrality, indicating a significant role in bridging diverse music styles and artist communities. High betweenness scores for artists like Romeo Santos and Drake suggest they play critical roles in connecting disparate parts of the network, likely bridging international and genre-specific gaps.

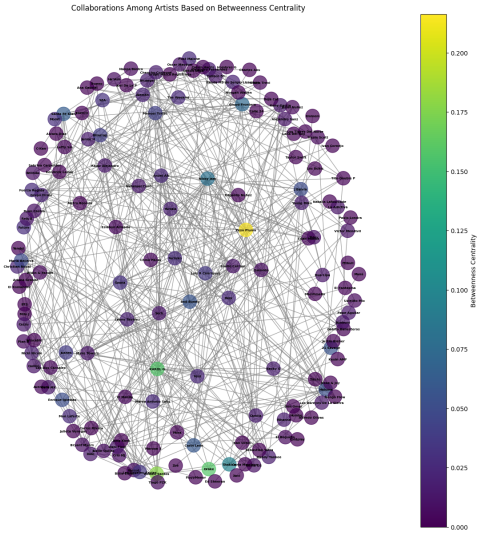

Fig. 11. Collaborations Among Artists Based on Betweenness Centrality

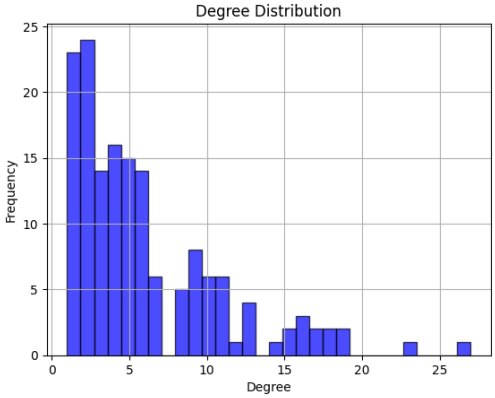

Fig. 12. Degree Distribution

5) Degree Distribution: The degree distribution shows a somewhat bell-shaped curve skewed towards lower degrees, indicating that while many artists engage in collaborations, a few highly connected individuals act as central hubs. This pattern supports a network where influential artists help disseminate new music and trends across the scene, potentially making the network resilient and dynamic.

Understanding these network dynamics can aid industry stakeholders in identifying key artists for promotional campaigns, strategic collaborations, and market expansion efforts. Artists with high betweenness centrality are particularly valuable for campaigns aiming to bridge different musical communities.

IV. Link Prediction

Link prediction is a fundamental technique used in the analysis of social networks to forecast potential future connections between nodes. It involves estimating which new interactions (edges) are likely to occur between nodes based on existing network data and patterns. The primary goal of link prediction is to understand and predict the growth of the network by identifying the most probable future collaborations or connections between its members.

Given the complex and dynamic structure of networks, especially in contexts like the music industry, employing sophisticated algorithms for link prediction, such as those based on network structure as mentioned in, can significantly improve the accuracy and relevance of the predictions [3].

A. Techniques Used

1) Jaccard Coefficient: The Jaccard coefficient is a measure used in statistics to gauge the similarity and diversity of sample sets. In the context of network analysis, it is used as a metric for link prediction by evaluating the likelihood of a new link forming between two nodes based on their shared connections [4].

The formula is given by:

$$\sigma(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|}$$

In this context, the Jaccard Coefficient is particularly useful for predicting potential future collaborations. By measuring how many common collaborators two artists have and comparing this to their total distinct collaborators, the Jaccard Coefficient can highlight artists who are more likely to collaborate due to shared connections. This can be particularly insightful for music producers or event organizers looking to create new music collaborations or lineups for music festivals.

For platforms like Spotify, using the Jaccard Coefficient to predict which artists might collaborate in the future can enhance user recommendations, not just for existing songs but for potential new releases and artist discoveries.

2) Resource Allocation and Community Resource Allocation: Resource allocation is a link prediction measure that simulates how resources (information, influence, etc.) could flow through a network based on the shared connections of two nodes [4]. The formula is given as:

$$\sigma(v_i, v_j) = \sum_{v_k \in N(v_i) \cap N(v_j)} \frac{1}{|N(v_k)|}$$

The intuition behind this is that if two nodes share a common neighbor with fewer connections, there is a stronger potential for a link to form between them because the shared neighbor might be a more focused channel of resource transfer or influence.

On the other hand, community resource allocation extends the basic concept of resource allocation by focusing only on common neighbors within the same community as the nodes being considered. This is particularly relevant in networks where community structure is prominent, and interactions are more frequent within the same community. The modified formula is the following:

$$\sigma(v_i, v_j) = \sum_{v_k \in N(v_i) \cap N(v_j)} f(v_k) \frac{1}{|N(v_k)|}$$

where $f(v_k)$ is defined as:

$$f(v_k) = \begin{cases} 1 & \text{if } v_k \text{ is in the same community as } v_i \text{ and } v_j \\ 0 & \text{otherwise} \end{cases}$$

Music industry professionals can use these measures for strategic planning, such as in festival lineups that could be likely to attract large audiences by featuring artists who are predicted to collaborate or share a common fan base; or in A&R (artist and repertoire) strategies, where managers can identify emerging artists who are likely to collaborate with established artists, providing opportunities for mentorship and growth within the community.

B. My Top Artists

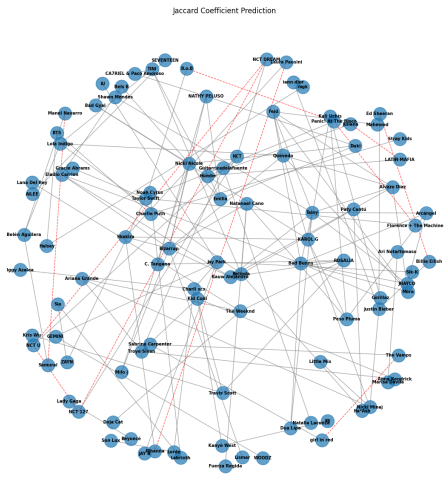

Fig. 13. Jaccard Coefficient Prediction

Based on Jaccard Coefficient, most probable links based on my top artists where:

- Rihanna and Laura Pausini
- NCT U and NCT DREAM
- NCT U and NCT 127
- Juliana and LATIN MAFIA
- The Vamps and girl in red
- Samuraï and Manel Navarro
- Billie Eilish and Ed Sheeran
- NCT DREAM and NCT 127
- Panic! At The Disco and Florence + The Machine
- Panic! At The Disco and B.o.B

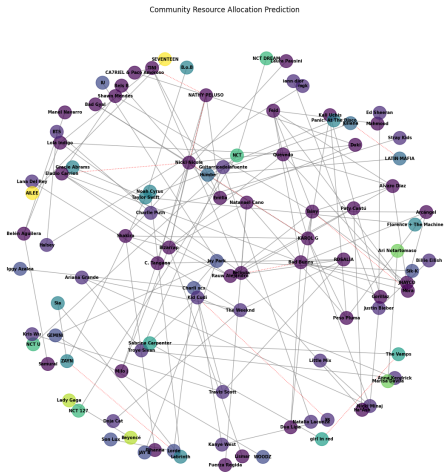

Fig. 14. Community Resource Allocation Prediction

In the case of resource allocation and community resource allocation, the results were the same, with the most probable links being between:

- TINI and NATHY PELUSO, with a score of 0.75
- Mora and Tainy, with 0.73
- Nicki Nicole and NATHY PELUSO, with 0.69
- Nicki Nicole and Eladio Carrion, with 0.61
- Rauw Alejandro anad Bad Bunny, with 0.58
- Labrinth and ZAYN, with 0.5
- KAROL G and Nicki Nicole, with 0.5
- Kid Cudi and ¥$, with 0.5
- Juliana and LATIN MAFIA, with 0.5
- The Vamps and girl in red, with 0.5

C. Mexico's Top Artists

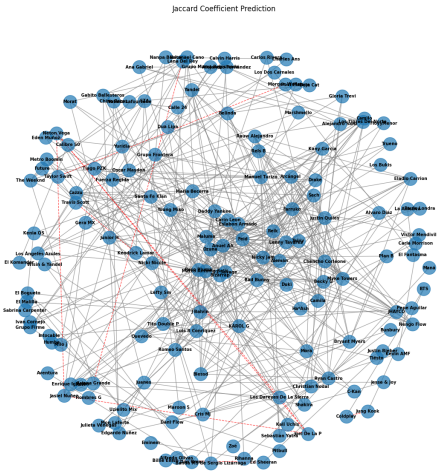

Fig. 15. Jaccard Coefficient Prediction

Based on Jaccard Coefficient, most probable links where:

- Billie Eilish and Ed Sheeran
- Taylor Swift and Morgan Wallen
- Ariana Grande and Lana Del Rey
- Joel De La P and Neton Vega
- Coldplay and Jung Kook
- Joel De La P and Jasiel Nuñez
- Joel De La P and Los Dareyes De La Sierra
- Jasiel Nuñez and Neton Vega
- Los Dareyes De La Sierra and Neton Vega
- Kendrick Lamar and Travis Scott

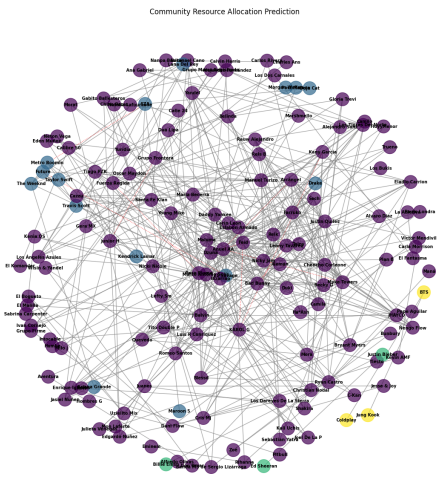

Fig. 16. Community Resource Allocation Prediction

In the case of resource allocation and community resource allocation, it happened the same as with my top artist networks and the results were the same, with the most probable links being between:

- Myke Towers and Anuel AA, with 0.90
- Future and 21 Savage, with 0.77
- Junior H and Grupo Frontera, with 0.74
- Feid and Anuel AA, with 0.73
- Metro Boomin and SZA, with 0.7
- Lenny Tavárez and Bad Bunny, with 0.68
- KAROL G and Kany García, with 0.66
- Peso Pluma and Fuerza Regida, with 0.66
- Anuel AA and Arcángel, 0.63
- KAROL G and Feid, with 0.63

V. Conclusions

This analysis of social networks, centered on artist collaborations, reveals intricate patterns and interactions among artists within two distinct networks: my top artists and Mexico's top artists. Utilizing advanced network analysis techniques, the study successfully identified existing collaboration patterns and accurately predicted potential future collaborations.

The findings from the mapping and visualization phases underscore the diversity and complexity of the musical landscapes in both personal and regional contexts. The analysis observed moderate connectivity within my top artist network and a highly interconnected community within Mexico's top artist network, reflecting the vibrant collaborative environment in the Mexican music scene.

The application of link prediction methods, such as the Jaccard coefficient and resource allocation, has further enhanced the understanding of how artists are likely to collaborate in the future, offering valuable insights for strategic decision-making in the music industry. By quantifying the relationships between artists and predicting future collaborations, this analysis not only aids in understanding current trends but also in anticipating shifts in the musical landscape.

Ultimately, this project shows the power of network analysis in the digital age, offering a robust tool for artists, producers, and industry stakeholders to navigate the increasingly complex world of music production and distribution. Industry professionals can leverage the insights garnered here to forge strategic partnerships, curate engaging musical events, and drive innovation in the ever-evolving music industry.

References

[1] Spotify, "Web API," Spotify for Developers. https://developer.spotify.com/documentation/web-api (accessed Aug. 03, 2024).

[2] P. Lamere, "Spotipy 2.0 Documentation," Spotipy, 2014. https://spotipy.readthedocs.io/en/2.24.0/ (accessed Aug. 03, 2024).

[3] S. Bhattacharya, S. Sinha, P. Dey, A. Saha, C. Chowdhury, and S. Roy, "Online social-network sensing models," in Elsevier eBooks, 2023, pp. 113–140. doi: 10.1016/b978-0-32-390535-0.00010-0.

[4] D. O. Gamboa Angulo, "Link Prediction," Jul. 25, 2022. https://upy-my.sharepoint.com/:p:/r/personal/didier_gamboa_upy_edu_mx/_layouts/15/Doc.aspx?sourcedoc=%7B353BF3FB-681F-4795-89B4-50FED617C552%7D&file=L3.1_LinkPrediction.pptx&action=edit&mobileredirect=true