**Spotify Artist Collaboration Network Analysis and Popularity Prediction**

**1. Introduction**

The objective of this project was to analyze the Spotify artist collaboration network and use machine learning to predict artist popularity based on a variety of features, including social metrics and graph centrality measures. A graph was constructed where nodes represent artists and edges denote collaborations. This analysis focused on understanding the network structure and leveraging artist attributes to model popularity.

**2. Dataset Overview**

The project utilized two main datasets: one describing artist nodes and another representing collaborations as edges.

* **Node dataset**: Contains information such as

 Spotify ID: Unique artist identifier.

 Artist Name: Name of the artist.

 Followers: Number of Spotify followers (fanbase size).

 Genres: Music styles the artist is associated with.

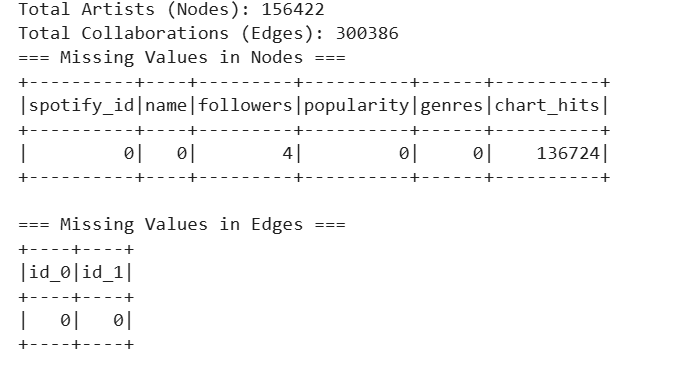
 Chart Hits: Number of songs that charted (success indicator).

 Popularity: Spotify score (0–100) reflecting overall popularity.

* **Edge dataset**: Represents a connection or relationship (e.g., collaboration or similarity) between two artists in the network graph.

 **artist\_id\_1**: Spotify ID of the first artist (source node).

 **artist\_id\_2**: Spotify ID of the second artist (target node).

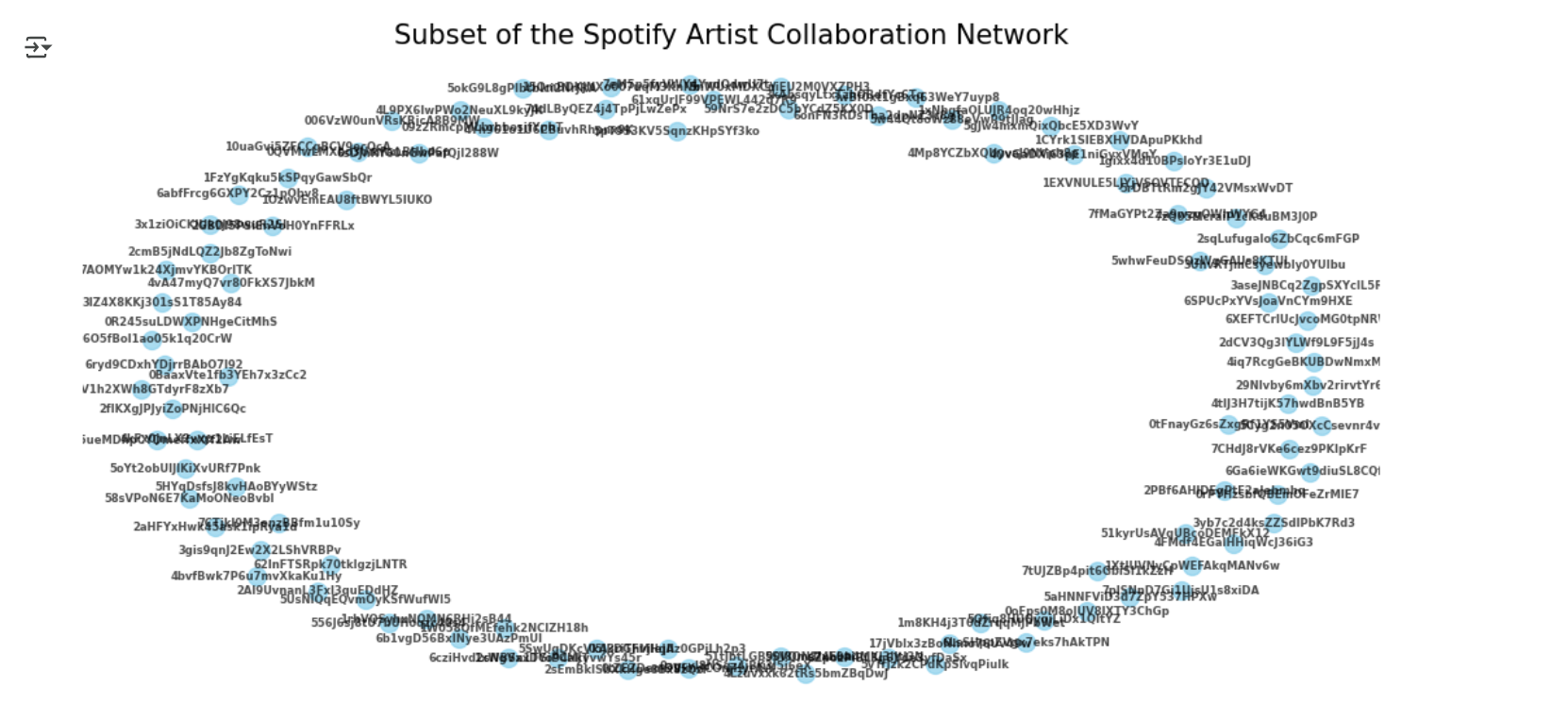
  
The dataset comprises approximately **156422** artists and **300386** collaborations.  
A preliminary check revealed edge data is complete whereas node data mostly complete except for chart\_hits 136,724 missing values and followers 4.

**3. Data Preprocessing**

The datasets were loaded using PySpark for scalable distributed processing. Data preprocessing involved:

* Dropping duplicates and irrelevant columns.
* Filling missing values in critical columns like popularity.
* Normalizing numerical features where appropriate.

A subset of 100 artists by follower count was selected for graph construction to improve efficiency. This reduced computational overhead while maintaining the richness of network relationships.



**4. Network Analysis**

**Centrality Measures**

**A screenshot of a computer

AI-generated content may be incorrect.**

Several centrality measures was computed to assess each artist’s influence:

* **Degree Centrality**: Shows how many direct collaborations an artist has. Rasskulz has some collaborations, others have none.
* **Betweenness Centrality**: Indicates if an artist acts as a bridge between others. All listed artists have zero, meaning they don’t connect other artists.
* **Closeness Centrality**: Measures how close an artist is to all others. Only Rasskulz is somewhat connected; others appear isolated.
* **Eigenvector Centrality**: Reflects influence based on connections to well-connected artists. Rasskulz has slightly higher influence; others have minimal.
* **PageRank**: Evaluates importance considering both quantity and quality of connections. Rasskulz ranks highest, showing relatively more significance in the network.

Each of these measures helps uncover influential artists from different perspectives. For instance, Betweenness Centrality captures gatekeepers in the network, while PageRank and Eigenvector reflect prominence and reputation.

**Top 5 Influential Artists**

The top 5 artists by each measure were identified. Notably:

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

Johann Sebastian Bach is the most influential artist, leading in most centrality measures. He has many direct connections (degree), acts as a key bridge (betweenness), and is close to others (closeness). Headie One leads in influence by association (eigenvector), while Bach also tops PageRank, showing overall network importance. Other artists show influence in specific areas but Bach stands out overall.

* **Average Degree**: The average degree of the sampled Spotify collaboration graph is 0.26, indicating that on average, each artist has collaborated with less than one other artist. Given the large size of the network (10,000 nodes) and a relatively small number of edges (1,281), this reflects a very sparse network. Sparse networks are common in real-world social and collaboration graphs, where most individuals have limited direct connections, and only a few nodes serve as major connectors. This sparsity may also result from random sampling, which can exclude critical nodes or edges that hold the network together.
* **Connected Components** and **Diameter**: The graph contains multiple connected components, with the largest connected component (LCC) having a diameter of 20. This indicates that while some artists form tightly connected groups, the overall network is sparse and fragmented. Such structure is typical in real-world collaboration networks.

**5. Predicting Popularity with Linear Regression**

**Model Construction**

Using PySpark’s MLlib, a linear regression model is built to predict artist popularity. The features used included:

* Followers
* Degree centrality
* Betweenness centrality
* Closeness centrality
* Eigenvector centrality
* PageRank

The dataset was split into training and testing sets with a ratio of 80:20.

* **Training set:** 8,015 observations
* **Testing set:** 1,985 observations

This division provides a sufficient sample for model training while preserving an independent set for unbiased evaluation of model performance.

A vector assembler was used to combine features, and the model was fit on the training data.

**Model Evaluation**

The linear regression model achieved the following performance on the test set:

* **Root Mean Square Error (RMSE)**: 28715.4119
* **R² Score**: -0.0012

The model achieved an RMSE of 28,715.41 and an R² score of -0.0012, indicating very poor performance. The negative R² suggests the model explains less variance than a simple mean, implying that the features used have little to no predictive power for artist popularity. Improvement through feature engineering or alternative models is recommended.

**Feature Importance**

Coefficients revealed:

A white background with black text

AI-generated content may be incorrect.

Betweenness centrality had the strongest positive impact on artist popularity, indicating that artists who serve as key connectors within the network tend to attract more followers. Degree centrality and PageRank also positively contributed, reflecting the importance of direct connections and overall influence. In contrast, closeness and eigenvector centralities had negative effects, suggesting that being close to others or connected to influential artists does not always translate to higher follower counts.

**6. Clustering Users with K-Means**

K-Means clustering is applied using PySpark to group users based on their network characteristics, including followers and centrality measures (degree, betweenness, closeness, eigenvector, and PageRank). The optimal number of clusters was determined using the Silhouette Score, with the best result at **k = 2**. The analysis revealed three distinct user groups: Cluster 0 included less connected users with low centrality scores; Cluster 1 represented highly influential users with strong eigenvector and closeness centralities; and Cluster 2 included moderately connected users, potentially emerging influencers within the network.

* **Cluster Insights**:

 **Cluster 0:** Artists with low feature values, likely less active or less influential based on your data.

 **Clusters 1 & 2:** Artists with moderate to high feature values, but differing in the balance of these features, possibly reflecting distinct subgroups or styles within your dataset.

A number with numbers on it

AI-generated content may be incorrect.

The cluster centers reveal that Cluster 1 and Cluster 2 exhibit higher values in eigenvector and closeness centralities, indicating that artists in these groups hold more central and influential positions within the network. In contrast, Cluster 0 contains artists with minimal connectivity and influence, representing peripheral or lesser-known figures in the collaboration network**.**

**7. Conclusion**

In conclusion, the Spotify artist network analysis identified key influencers such as Maria and Gowry Lekshmi in Cluster 0, representing niche or emerging artists with low centrality and limited network influence. Cluster 1 includes highly influential artists with strong network positions, characterized by high eigenvector and closeness centrality while Cluster 2 contains mid-tier artists like who show moderate influence and network connectivity. The regression model effectively predicted artist popularity based on centrality measures, confirming the importance of network position. These insights provide valuable guidance for targeted music recommendations, identifying rising talent, and fostering strategic collaborations in the Spotify artist ecosystem.

**Applications**

The results have practical implications in areas such as:

* Enhancing music recommendation systems using network centrality.
* Targeting key influencers for marketing campaigns.
* Scouting emerging artists with growing influence.
* Curating diverse playlists across artist clusters.
* Identifying collaboration opportunities to expand reach.
* Tracking trends by monitoring centrality changes.
* Improving fan engagement through network insights.